Effective and Practical Strategies for Combatting Misinformation

Catherine King

CMU-S3D-25-102

May 2025

Software and Societal Systems Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

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

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

This work was supported by the Knight Foundation, the Hillman Family Foundation, the Office of Naval Research (ONR) MURI grants N000142112749/N000141712675, ONR Minerva grants N000141310835/N000141612324, ONR Bot Hunter grant N000141812108, and PARALLAX FA8650-15-D-658. Additional support was provided by Carnegie Mellon University's Graduate small Project Help (GuSH), 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 author and should not be interpreted as representing the official policies, either expressed or implied, of the Knight Foundation, the Hillman Family Foundation, ONR, PARALLEX, or the U.S. government.

Keywords: misinformation interventions, countermeasures, survey analysis, public opinon, social network analysis, social media, social cybersecurity, computational social science

To my family

Abstract

Social media platforms, which are becoming a primary news source for many individuals, can quickly spread misinformation faster than ever before. These information disorders may contribute to increased polarization and extremism, possibly undermining 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 evaluated without complete access to comprehensive social media data. Most researchers also focus on assessing the effectiveness of an intervention without considering whether 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 created a detailed categorization of interventions. Then, I conducted a citation network analysis of the literature in this field to identify research gaps. I administered a comprehensive survey asking the American public about their social media behavior and opinions on various interventions. The survey also examines how certain factors may influence user acceptance, including the perceived effectiveness, fairness, and intrusiveness of each intervention. Next, I developed a training effort to assess whether media literacy can improve an individual's willingness and ability to counter misinformation. Finally, I combine this research with the professional opinions of expert researchers in this field to evaluate countermeasures, aiming to identify the shared features that make interventions both effective and practical.

Acknowledgments

Completing a PhD is no easy task, even in the best circumstances, and I started mine at the beginning of the pandemic. I could not have reached my defense without the supportive community around me and several incredible individuals who helped me get through this challenging yet rewarding time.

First, I would like to thank my advisor, Kathleen Carley, for taking a chance on me and guiding me on this journey. I feel incredibly fortunate to have worked and collaborated with her, as well as many other brilliant researchers in her research group, and to have pursued my intellectual interests so freely. I am also very grateful to my committee members, Hong Shen, Chris Labash, and Pablo Barberá, for providing thoughtful feedback that pushed my work to be the best it could be. Thank you all for helping me become a better researcher and academic.

I wouldn't have pursued a PhD without the enthusiastic support of my mentors at William and Mary. Thank you to Ryan Vinroot for being such a supportive undergraduate advisor and for encouraging me to consider research. I also want to express my gratitude to my undergraduate research mentors, Sarah Day and Drew LaMar, and my master's research mentor, Lawrence Leemis. Thank you for helping me develop my research skills and for encouraging me to apply to graduate school.

I have been very lucky to have worked with such excellent researchers in CASOS during my time here. Thank you to Christine Lepird for being a great research partner on the OMEN project, Peter Carragher for being a great collaborator on my literature review, and Samantha Phillips for helping me design and deploy the large-scale public opinion survey. I am lucky to have co-authored papers with the other members of my cohort, Daniele Bellutta and J.D. Moffitt. I would also like to thank Charity Jacobs for being a wonderful TA partner, Matt Hicks for his work on the scenario generator and his expertise on ChatGPT prompting, and my former OMEN teammates, including Janice Blane and Geoff Dobson. Finally, I would like to thank Sienna Watkins, who goes above and beyond for all of us. She truly is superhuman.

My PhD experience was enriched by the great friends I made along the way. Maria Casimiro and I became fast friends, and our Zoom chats during the pandemic were a lifeline. Janice facilitated making the CASOS group more social and collaborative during her time here, and Christine and I have enjoyed many movie nights together. I want to thank Peter for teaching me (and many others) how to play Dungeons and Dragons, which has been a fun and necessary escape during some of the more stressful moments. I have greatly enjoyed playing DnD with so many people in S3D, including Evan Williams, Eli Claggett, Carolina Carreira, Stephen Dipple, Nikitha Rao, Samantha Phillips, and Vasu Vikram.

Finally, I would like to thank my family for their tireless support. My parents have always encouraged me to work hard and pursue higher degrees. My sisters have always been there for me, no matter what. And finally, to my beautiful and loving husband, Travis. I genuinely don't have the words to express my gratitude for your love and support throughout the years. I couldn't have done it without you.

Contents

1 Introduction 1
1.1 Thesis Objective . . . 1
1.2 Background . . . 2
1.2.1 Social Cybersecurity . . . 2
1.2.2 Definitions . . . 2
1.3 Misinformation Categorization . . . 3
1.3.1 Literature Review . . . 3
1.3.2 Misinformation Typology . . . 4
1.3.3 Misinformation Pipeline . . . 8
1.4 Countermeasures Categorization . . . 11
1.4.1 Literature Review . . . 11
1.4.2 Countermeasures Typology . . . 13
1.4.3 Targeting the Misinformation Pipeline . . . 14
1.5 Platform Policies . . . 15
1.5.1 Community Guidelines . . . 15
1.5.2 Enforcement . . . 17
1.6 Data . . . 17
1.7 Chapter Overview . . . 19

2 Bibliometric Analysis of the Interventions Literature 21
2.1 Introduction . . . 21
2.2 Data and Methods . . . 22
2.2.1 Paper Labels . . . 22
2.2.2 Inclusion Criteria . . . 24
2.2.3 Literature Search . . . 24
2.2.4 Training of Labelers . . . 25
2.2.5 ChatGPT Prompts . . . 27
2.2.6 Label Agreement . . . 29
2.2.7 Data Availability . . . 30
2.3 Citation Network Analysis . . . 30
2.3.1 Data Attributes . . . 30
2.3.2 Nodesets . . . 31
2.3.3 Networks . . . 31
2.4 Results . . . 32

2.4.1 Descriptive Statistics 32
2.4.2 Publication Venue Analysis 32
2.4.3 Topic Analysis 37
2.4.4 Author Analysis 37
2.5 Discussion 42
2.6 Conclusions 44
2.6.1 Limitations 44
2.6.2 Contributions 44

3 Characterizing User-based Countermeasures 45
3.1 Introduction 45
3.2 Data and Methods 47
3.2.1 Ethics Information 47
3.2.2 Pilot Data 47
3.2.3 Survey Design and Sampling Plan 47
3.2.4 Measures 49
3.3 Closeness Analysis 50
3.3.1 Related Work 51
3.3.2 Bayesian Power Analysis 52
3.3.3 Statistical Methods 56
3.3.4 Results 57
3.3.5 Closeness Discussion 66
3.4 Platform Analysis 67
3.4.1 Related Work 67
3.4.2 Results 73
3.4.3 Platform Discussion 77
3.5 Overall Discussion 78
3.6 Conclusions 79
3.6.1 Limitations 79
3.6.2 Contributions 80

4 Improving User-based Countermeasures 81
4.1 Introduction 81
4.2 Related Work 82
4.2.1 Media Literacy Interventions 82
4.2.2 User-Based Interventions 82
4.3 Data and Methods 83
4.3.1 OMEN Overview 83
4.3.2 Participant Demographics 83
4.3.3 Survey Design 84
4.3.4 Training Sessions 90
4.3.5 Analysis Methods 92
4.3.6 Ethics Information 93
4.4 Results: Misinformation Detection 93

4.4.1 Accuracy 93
4.4.2 Confidence 95
4.5 Results: Countering Misinformation 96
4.5.1 Selected Interventions 96
4.5.2 Effort Level 97
4.5.3 Platform and Poster Impact 97
4.5.4 Factors Affecting Countering Actions 99
4.6 Participant Feedback 103
4.7 Discussion 105
4.8 Conclusions 106
4.8.1 Limitations 106
4.8.2 Contributions 106

5 Characterizing Platform and Government Countermeasures 109
5.1 Introduction 109
5.2 Related Work 110
5.2.1 Intervention Selection 111
5.3 Data and Methods 113
5.3.1 Platform and Government Survey Questions 113
5.3.2 Measures 114
5.3.3 Pre-Registered Analysis Plan 114
5.3.4 Adhoc Analysis 115
5.4 Results 115
5.4.1 Factors that Influence Support for Interventions 115
5.4.2 Overall Support and Perceptions of Interventions 115
5.4.3 Individual Differences in Support and Perceptions of Interventions 117
5.4.4 Partisanship Influences Features that Predict Support 123
5.5 Discussion 123
5.6 Conclusions 126
5.6.1 Limitations 126
5.6.2 Contributions 126

6 Recommendations for Effective and Practical Countermeasures 129
6.1 Introduction 129
6.2 Related Work 130
6.2.1 Operationalized Interventions 130
6.2.2 Objective Characteristics 130
6.2.3 User Perceptions 133
6.2.4 Main Criteria 135
6.3 Data and Methods 136
6.3.1 Survey Design 136
6.3.2 Participant Demographics 138
6.3.3 Analysis Methods 139
6.3.4 Measures 139

6.4 Results and Discussion 139
6.4.1 Content Distribution 142
6.4.2 Content Moderation 144
6.4.3 Account Moderation 146
6.4.4 Content Labeling 147
6.4.5 User-based Measures 149
6.4.6 Media Literacy and Education 150
6.4.7 Institutional Measures 151
6.4.8 Generative AI 153
6.5 Top-Ranked Interventions 154
6.6 Conclusions 156
6.6.1 Limitations 156
6.6.2 Contributions 157

7 Concluding Remarks 159
7.1 Contributions 159
7.1.1 Theoretical Contributions 159
7.1.2 Empirical Contributions 160
7.2 Summary of Key Findings 161
7.3 General Implications 162
7.3.1 The Hypocrisy Gap 162
7.3.2 Pipeline Coverage and Intervention Prioritization 163
7.3.3 Policy Enforcement vs. Perceptions 164
7.3.4 Potential Backfire Conditions 166
7.3.5 Emergence of Generative AI 166
7.4 Recommendations 167
7.4.1 For Users 167
7.4.2 For Platforms 168
7.4.3 For Policymakers 169
7.4.4 For Institutions 170
7.5 Limitations and Future Research 171
7.6 Final Comments 172

A Countermeasures Categorization 173
A.1 Content Distribution 173
A.2 Content Moderation 174
A.3 Account Moderation 174
A.4 Content Labeling 175
A.5 User-based Measures 175
A.6 Media Literacy and Education 176
A.7 Institutional Measures 176
A.8 Generative AI and Other 177

B Chapter 2: ChatGPT Prompts 179
B.1 Effectiveness Labeling Prompt . . . 179
B.2 Acceptance Labeling Prompt . . . 180
B.3 Interventions Labeling Prompt . . . 180

C Chapter 3: Pre-Registered Hypotheses and Analyses 183

D Chapter 3: Supplemental Categorical Analysis 187

E Chapter 3: Demographic Analysis 191

F Chapter 4: OMEN Survey Posts 195
F.1 Misinformation Detection Posts . . . 195
F.2 Countering Posts . . . 211

G Chapter 5: Robustness Analysis 219

H Chapter 5: Demographic Analysis 223
H.1 Intervention Ratings by Partisanship . . . 223
H.2 One-way ANOVA Statistical Tests . . . 229

Bibliography 231

List of Figures

2.1 Literature review process. 25

2.2 Number of articles in this review published per year. 34

2.3 Co-Publication Venue network. Nodes are sized by total degree centrality and colored by discipline (red for computer science, yellow for psychology, blue for social sciences, black for multidisciplinary, and grey for other.) 36

2.4 Number of articles in this review by venue discipline. 36

2.5 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. 39

2.6 Number of articles in this review by general intervention category studied. 40

2.7 The largest component of the Co-Authorship network. Nodes are sized by total number of papers and colored by betweenness centrality, with lighter blue colors indicating higher betweenness. 41

2.8 The largest component of the Co-Authorship network excluding one paper with 30 authors. Nodes are sized by total number of papers and colored by betweenness centrality, with lighter blue colors indicating higher betweenness. 42

3.1 Total number of participants who selected each intervention type from Table 3.1 at least once in their behavioral and opinion responses. For comparison purposes, only participants who had seen misinformation at that closeness level have their opinion counts included. 59

3.2 Number of participants who reported expending a maximum of no, low, or high effort when seeing misinformation compared with the total number of those same participants who ***believe*** one should expend no, low, or high effort when seeing misinformation at each of the three closeness levels and against oneself. 60

3.3 Total number of participants who had posted misinformation and selected each intervention type from Table 3.2 at least once in their behavioral and opinion responses. For comparison purposes, only participants who admitted to posting misinformation had their opinion counts included. 61

3.4 Highest effort level participants said one should respond with when seeing misinformation posted by others or oneself broken up by party, age, gender, race, education, and income. Chi-sq tests: $p < 0.05$*, $p < 0.01$**, and $p < 0.001$***. . 65

3.5 The number of participants who use each platform at least weekly or daily compared with the number of participants who have seen misinformation on that platform. 73

3.6 Highest effort level participants said they responded with when seeing misinformation posted by others on the top 6 platforms for misinformation. Chi-sq test: $p < 0.001$***. 75

3.7 Highest effort level participants said they responded with when seeing misinformation posted by **not close contacts only** on the top 6 platforms for misinformation. Chi-sq test: $p < 0.001$***. 76

3.8 Number of participants who admitted to posting misinformation on each platform. 76

3.9 Highest effort level participants said they responded with after realizing they had posted misinformation accidentally among the top 4 platforms. Chi-sq test was not significant. 77

4.1 An example of a false post shown to participants. 88

4.2 Summary training slide on misinformation detection. 90

4.3 The number of times participants said their answer would change depending on either the platform or the poster over both the pre- and post-training surveys. . . . 96

4.4 The number of times participants said their answer would change depending on either the platform or the poster in both the pre and post-training surveys. 98

4.5 The number of times participants said their answer would change depending on either the poster or the platform combined over both the pre- and post-training surveys. 99

4.6 This figure displays the total number of times over both surveys that participants mentioned a specific sub-theme when asked if and how their answer would change depending on the misinformation poster and the platform. 102

4.7 This figure summarizes the responses to the feedback questions at the end of the study, indicating whether participants agreed or disagreed that the training was helpful. 104

4.8 The number of participants who said they used each technique in the post-training survey. 105

5.1 Estimate and 95% CI of the effect of perceived fairness, effectiveness and intrusiveness on support depending on implementer. 117

5.2 Average and 95% CI support, perceived fairness, perceived effectiveness, and perceived intrusiveness for each intervention (1-10) and implementer (government and platform) on a 1-5 Likert scale. 118

5.3 Average ratings by political party and gender. One-way ANOVA tests were run on each grouping, with stars indicating the level of significance: $p < 0.05$*, $p < 0.01$**, and $p < 0.001$*** 119

5.4 Average ratings by age, education, income, political ideology, and misinformation exposure. One-way ANOVA tests were run on each grouping, with stars indicating the level of significance: $p < 0.05$*, $p < 0.01$**, and $p < 0.001$*** . . 121

5.5 Average Likert ratings and 95% confidence intervals by political party and gender, separated by intervention implementer. 122

6.1 Stacked bar plot representing the percentage of responses on the effectiveness and acceptance of the general categories of interventions. 140
6.2 Average effectiveness and acceptance Likert scores per general intervention category. 141
6.3 A heatmap showing the average score for each metric across general intervention categories. 156

F.1 Pre-Test Misinformation Post #1. 195
F.2 Pre-Test Misinformation Post #2. 196
F.3 Pre-Test Misinformation Post #3. 197
F.4 Pre-Test Misinformation Post #4. 198
F.5 Pre-Test Real News Post #1. 199
F.6 Pre-Test Real News Post #2. 200
F.7 Pre-Test Real News Post #3. 201
F.8 Pre-Test Real News Post #4. 202
F.9 Post-Test Misinformation Post #1. 203
F.10 Post-Test Misinformation Post #2. 204
F.11 Post-Test Misinformation Post #3. 205
F.12 Post-Test Misinformation Post #4. 206
F.13 Post-Test Real News Post #1. 207
F.14 Post-Test Real News Post #2. 208
F.15 Post-Test Real News Post #3. 209
F.16 Post-Test Real News Post #4. 210
F.17 Pre-Test Countering Post #1. 211
F.18 Pre-Test Countering Post #2. 212
F.19 Pre-Test Countering Post #3. 213
F.20 Pre-Test Countering Post #4. 214
F.21 Post-Test Countering Post #1. 215
F.22 Post-Test Countering Post #2. 216
F.23 Post-Test Countering Post #3. 217
F.24 Post-Test Countering Post #4. 218

H.1 Average Likert ratings for the **friction** intervention: temporarily delaying users from posting content they have not opened. 224
H.2 Average Likert ratings for the **advertising policy** intervention: requiring all ads to be put through a fact-checking process. 225
H.3 Average Likert ratings for the **algorithmic downranking** intervention: de-emphasizing posts that are verified to contain misinformation. 225
H.4 Average Likert ratings for the **content removal** intervention: removing posts verified to contain misinformation. 226
H.5 Average Likert ratings for the **account removal** intervention: permanently banning users who post misinformation a certain number of times. 226
H.6 Average Likert ratings for the **misinformation disclosure** intervention: notifying users if they posted content verified to contain misinformation. 227

H.7 Average Likert ratings for the **fact-check labelling** intervention: publicly labelling posts verified to contain misinformation with information about and from verified sources. 227
H.8 Average Likert ratings for the **media literacy** intervention: investing in digital media literacy and promoting educational content about detecting misinformation on and offline. 228
H.9 Average Likert ratings for the **media support** intervention: promoting and investing in local media, which is thought to be most in tune with local norms, culture, and context. 228
H.10 Average Likert ratings for the **data sharing** intervention: regularly releasing data and/or internal research reports about misinformation to the public and outside researchers. 229

List of Tables

1.1 Misinformation categorizations in the literature. 3
1.2 Proposed misinformation typology. 8
1.3 Proposed misinformation pipeline. 11
1.4 Categorizations of misinformation interventions in the literature. 12
1.5 Proposed categorization of misinformation interventions. 14
1.6 The categories of countermeasures and the phases of the misinformation pipeline they target. 15
1.7 Current platform policies regarding misinformation. 16
1.8 Summary of the datasets used in this thesis. 18

2.1 Intervention topic labels. 23
2.2 The intervention labels used the least after Stage 1, sorted from highest to lowest according to the number of papers assigned to those labels. 26
2.3 Summary of interventions labeling by different prompt and ChatGPT model types. The values indicate Jaccard similarity scores that compare the labeled output produced by the model run against the final label. 28
2.4 Jaccard Similarity agreement with the final labels broken up by rater type. Rater types are sorted by total number of papers labeled. 29
2.5 Jaccard Similarity agreement with the final labels broken up by specific raters. Raters are sorted by total number of papers labeled. 30
2.6 The top 10 most cited papers in the data set. 33
2.7 The top 10 venues in the Co-Publication Venue network. 34
2.8 The top venues by total number of publications in the data set. 35
2.9 The number of papers and unique authors who have studied each intervention type. 38
2.10 Statistical summary of the number of papers and unique authors by intervention type. Outliers are identified from the inter-quartile range. 39
2.11 The number of studies examining the effectiveness or acceptance of one or more interventions. 39
2.12 The top 10 most cited authors with more than one paper in the data set. 40
2.13 The top 10 authors in the Co-Authorship network. 41

3.1 Actions social media users can take when they see misinformation online. 49
3.2 Actions social media users can take when they realize they have posted misinformation online. 49
3.3 Level of misinformation exposure among survey participants. 57

3.4 Pre-registered Bayesian paired hypothesis test results and interpretation for RQ1. Values in italics are for the hypotheses being run with the effort level summed as a robustness check. The p-value for the generalized McNemar's Chi-square Test is also included. 58

3.5 Pre-registered Bayesian paired hypothesis test results and interpretation for RQ2. Values in italics are for the hypotheses being run with the effort level summed as a robustness check. The p-value for the generalized McNemar's Chi-square Test is also included. 61

3.6 Pre-registered Bayesian paired hypothesis test results and interpretation for RQ3. Values in italics are for the hypotheses being run with the effort level summed as a robustness check. The p-value for the generalized McNemar's Chi-square Test is also included. 63

3.7 Social media platform categorization as developed by Zhu and Chen (2015) [322]. 69

3.8 Modified social media platform categorization. 69

3.9 Social media engagement on various platforms by dimension [296]. The innovaton and empowerment dimensions are excluded, as all platforms scored low. . . . 72

3.10 Number and percent of people who saw misinformation from close, somewhat, and not close contacts per platform. The top three platforms by percentage per closeness level are bolded. Platforms are sorted by the number of people who have seen misinformation on the platform. 74

3.11 Summary of interventions selected after posting misinformation. 77

4.1 The number of participants that used various social media platforms for work and for personal reasons. 85

4.2 Actions one can take when engaging with misinformation on social media. 87

4.3 The measured variables from Part 1 of the surveys. 93

4.4 The measured variables from Part 2 of the surveys. 94

4.5 The average and median number of questions answered correctly by participants. 95

4.6 Two-sided paired t-test results (df = 22) and the estimated effect size, Cohen's *d*, for the average accuracy scores given by each participant as defined in Table 4.3. 95

4.7 Two-sided paired t-test results (df = 22) and the estimated effect size, Cohen's *d*, for the average self-reported confidence scores on a scale 1-10 given by each participant as defined in Table 4.3. 96

4.8 The percentage and number of participants whose maximum effort level was in each of the three effort level categories described in Table 4.2. 97

4.9 The contingency table showing the number of participants that were no, low, and high effort in the pre vs. post-test. 97

4.10 Two-sided paired t-test results (df = 22) and the estimated effect size, Cohen's *d*, for the average poster and platform effect scores given by each participant as defined in Table 4.4. 98

4.11 The number of unique participants and total mentions of each major theme when considering the misinformation poster or the platform. 100

4.12 The number of unique participants that mentioned each sub-theme and the total number of times each sub-theme was mentioned when participants were asked to consider the poster of the misinformation. 100
4.13 The number of unique participants that mentioned each sub-theme and the total number of times each sub-theme was mentioned when participants were asked to consider the misinformation platform. 101
4.14 The average, standard deviation, and median feedback score on a 1-5 Likert scale, where 1 represents "strongly disagree" and 5 represents "strongly agree." . 104

5.1 Selected misinformation interventions. Text in [brackets] was included when the specified implementer was government. 112
5.2 OLS regression predicting support for a misinformation intervention as a function of perceived fairness, perceived effectiveness, perceived intrusiveness, and implementer (reference level: social media company) with robust standard errors clustered on participant and intervention. Model 1 is the main model. Model 2 is the robustness check that includes responses from all participants who responded to at least part of the survey. 116
5.3 OLS regressions predicting average support, perceived fairness, perceived effectiveness, and perceived intrusiveness of interventions as a function of age (18-24 to 65+ age brackets mapped to numeric values 0 to 5), gender (reference level: Female), partisanship (reference level: Democrat, combined "Other" party with "Independent"), political ideology (very liberal to very conservative categories mapped to numeric values 0 to 4), highest education level (less than high school diploma to Doctorate or Professional Degree categories mapped to numeric values 0 to 6), income (less than \$20,000 to over \$200,000 income brackets mapped to numeric values 0 to 7), frequency seeing misinformation (never to very often mapped to numeric values 0 to 4). 120
5.4 Contingency table of partisan and political ideology groups. Fisher's test comparing party and ideology categories ($\chi^2 = 1072, p < 0.001$). 122
5.5 Adhoc analysis: OLS regression predicting support for a misinformation intervention as a function of perceived fairness, perceived effectiveness, perceived intrusiveness, implementer (reference level: social media company), gender (reference level: Female), and partisanship (reference level: Democrat, with Independents combined with "Other") with robust standard errors clustered on participant and intervention. 124

6.1 Operationalized platform interventions. 131
6.2 Operationalized user-based and institutional interventions. 132
6.3 The percentage of respondents who agree or disagree that each general category of interventions is effective or acceptable to the public. 141
6.4 Objective characteristics of Content Distribution interventions. 142
6.5 Expert scores for Content Distribution interventions. Mean values are shown with their corresponding standard deviations in parentheses. 143
6.6 Objective characteristics of Content Moderation interventions. 144

6.7 Expert scores for Content Moderation interventions. Mean values are shown with their corresponding standard deviations in parentheses. 145
6.8 Objective characteristics of Account Moderation interventions. 146
6.9 Expert scores for Account Moderation interventions. Mean values are shown with their corresponding standard deviations in parentheses. 147
6.10 Objective characteristics of Content Labeling interventions. 147
6.11 Expert scores for Content Labeling interventions. Mean values are shown with their corresponding standard deviations in parentheses. 148
6.12 Objective characteristics of User-based interventions. 149
6.13 Expert scores for User-based Measures. Mean values are shown with their corresponding standard deviations in parentheses. 149
6.14 Objective characteristics of Media Literacy interventions. 150
6.15 Expert scores for Media Literacy and Education interventions. Mean values are shown with their corresponding standard deviations in parentheses. 150
6.16 Objective characteristics of Institutional Measures interventions. 151
6.17 Expert scores for Institutional Measures. Mean values are shown with their corresponding standard deviations in parentheses. 152
6.18 Objective characteristics of Generative AI interventions. 153
6.19 Expert scores for Generative AI interventions. Mean values are shown with their corresponding standard deviations in parentheses. 154
6.20 Ranking of the top 10 overall interventions by the five metrics. 155
6.21 The average metric values aggregated by intervention category. 155

7.1 Summary of key findings associated with user-based measures. 162
7.2 Summary of key findings associated with platform and government interventions. 162
7.3 Comparison of enforcement report features across five major social media platforms. 165

D.1 Contingency table of effort level, response type, and closeness level. Each cell contains the frequency of occurrences for each combination of the categorical variables. The chi-sq test statistic is shown: $p < 0.05^{*}$, $p < 0.01^{**}$, and $p < 0.001^{***}$. 187
D.2 The generalized estimating equation (GEE) model predicting the average frequency counts as a function of effort level (reference level: no effort), response type (reference level: behavior), and closeness (reference level: close contacts) when considering responses against others. 189
D.3 Contingency table of effort level and response type for actions against oneself. Each cell contains the frequency of occurrences for each combination of the categorical variables. The chi-sq test statistic is shown: $p < 0.05^{*}$, $p < 0.01^{**}$, and $p < 0.001^{***}$. 190
D.4 The generalized estimating equation (GEE) model predicting the average frequency counts as a function of effort level (reference level: no effort) and response type (reference level: behavior) when considering responses to oneself. . . 190

E.1 Demographics summary table. 191

G.1 OLS regressions predicting average support, perceived fairness, perceived effectiveness, and perceived intrusiveness of interventions as a function of demographic variables. Same analysis as Table 5.3, except partisanship separated out the "Other" party from "Independent". 220
G.2 OLS regressions predicting average support, perceived fairness, perceived effectiveness, and perceived intrusiveness of interventions as a function of demographic variables. Same analysis as Table 5.3, except partisanship was mapped to numeric values 0 to 4, with strong Democrats mapped to 0 and strong Republicans mapped to 4. 221
G.3 Adhoc analysis: OLS regression predicting support for a misinformation intervention as a function of perceived fairness, perceived effectiveness, perceived intrusiveness, implementer (reference level: social media company), gender (reference level: Female), and ideology (very liberal to very conservative categories mapped to numeric values 0 to 4) with robust standard errors clustered on participant and intervention. 222

H.1 Average support ratings and standard deviations per intervention by political party. The top 3 interventions by party are bolded 223
H.2 Top-ranked items by average support rating for each political affiliation. 224
H.3 F-statistics from one-way ANOVA tests comparing average ratings for support, perceived fairness, perceived effectiveness, and perceived intrusiveness across categories for each demographic variable. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$ 230

Chapter 1

Introduction

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 [174, 297, 320], and researchers are growing more concerned about how social media may be contributing to political polarization and to distrust in institutions and the media. Information disorders like misinformation and disinformation have been shown to have pressing societal impacts such as undermining democracy [287], increasing extremism [306], and lowering the uptake of various public health measures during a pandemic like COVID-19 [218].

While misinformation has existed long before social media, there are concerns that social media is exacerbating the issue by allowing it to spread faster than ever [297]. Over the past five years, Pew Research has found that about half of Americans get their news from social media at least sometimes [269], and nearly half report seeing misinformation daily on social media or elsewhere in their environment [195]. Additionally, most Americans believe that fake news causes confusion and contributes to extreme political views and hate crimes [3, 35]

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 [79]. Even if data is available, in some cases there are ethical challenges associated with sharing social media data with other researchers [44].

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 [234, 300], debunking [68, 96], and prebunking [180, 250]. However, many countermeasures, including those that could target creators of disinformation, have not been studied at all [79]. The lack of data access may contribute to why some interventions, like fact-checking, are studied significantly more than others. Finally, most studies focus on the effectiveness of the intervention without considering crucial aspects like user acceptance, political feasibility, and cost.

1.1 Thesis Objective

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 on acceptability while considering other relevant factors, such as fairness and cost. 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 create 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 potential government regulation, primarily focusing on the US-based context. Section 1.2 provides background context on the approach used throughout the dissertation and defines the general terms used. Sections 1.3 and 1.4 outline the various types of misinformation and countermeasures. Section 1.5 summarizes current platform policies. Finally, Section 1.6 discusses the data used throughout the dissertation, and Section 1.7 concludes with an outline of the subsequent chapters.

1.2 Background

1.2.1 Social Cybersecurity

Social media platforms have allowed people and organizations to access and spread information faster than ever before, and they have also facilitated the increased speed and reach of misinformation [174, 297, 320]. Researchers across various fields have been investigating the most effective and acceptable methods to combat fake news online. The research is in the emerging scientific area known as *social cybersecurity*, which investigates the impact of the online information space on society, culture, and politics [66, 209]. This research area analyzes information and network maneuvers, as well as their possible effects on human behavior and opinion.

This thesis employs an interdisciplinary approach put forth by this emerging research area, integrating social science theory and expertise with computational methods from network analysis and data science [66]. It utilizes computational social science techniques to identify and analyze the best ways to counter the proliferation of misinformation on social media platforms.

1.2.2 Definitions

Researchers typically make a distinction between misinformation and disinformation:

- **Misinformation** - False, inaccurate, and/or misleading information [273, 305, 313].
- **Disinformation** - Misinformation that is created with the *intent* to deceive and spread intentionally, often in the form of propaganda [273, 305, 313].

While this distinction is important, intent can be challenging to determine accurately. The focus of this thesis is not to detect mis/disinformation but rather to develop measures against false and misleading content. Any false, inaccurate, or misleading information, regardless of intent, will typically be referred to as "misinformation" for the remainder of this document.

1.3 Misinformation Categorization

This section first presents an overview and comparison of the ways different researchers classify misinformation. Then, building on this previous work, I outline a proposed typology.

1.3.1 Literature Review

This section examines four prominent review articles. Table 1.1 shows the high-level misinformation categories described in these documents.

First, Wardle et al. (2017) define seven main types of misinformation, loosely categorized along an axis that ranges from low risk and intent to deceive or harm to high risk and intent to deceive and harm. These categories, listed from low to high risk, are: *satire or parody*, *false connection*, *misleading content*, *false context*, *imposter content*, *manipulated content*, and *fabricated content* [305]. This typology has been adapted and used by several university library research guides, including Temple University [294], Northeastern University [289], and University of Iowa [80].

Table 1.1: Misinformation categorizations in the literature.

Type	**Wardle et al. (2017) [305]**	**Tandoc et al. (2017) [273]**	**Zannettou et al. (2019) [317]**	**Wang et al. (2020) [304]**
Satire and Parody	×	×	×	×
False Connection	×			
Misleading Content	×			
False Context	×			
Imposter Content	×			×
Manipulated Content	×	×		×
Fabricated Content	×	×	×	×
Error (False Content)				×
Propaganda		×	×	×
Conspiracy Theories			×	
Hoaxes			×	
Rumors			×	×
Advertising		×		×
Clickbait			×	×
Biased or one-sided			×	×
Other				×

Tandoc et al. (2017) review 34 articles that used the term “fake news” to develop a typology [273]. The authors primarily consider two dimensions: the level of facticity (the extent to which the misinformation uses facts) and the intention to deceive. They create six categories of fake news. Among the categories with low intentions to deceive are *news satire* and *news parody*. This article was the only one of the four that differentiated between satire and parody as separate categories. *News satire* typically employs exaggeration to deliver the news and relies on a high level of facticity (e.g., the Daily Show). In contrast, *news parody* is more likely to rely on fabricated content (e.g., the Onion) [273]. Among the categories with higher intentions to deceive, listed from high to low facticity, were misleading *advertising*, *propaganda*, *manipulated content*, and *fabricated content*. The authors note that advertising and propaganda may sometimes have overlapping motives [273].

Zannettou et al. (2019) define eight types of misinformation and detail the actors behind it and their potential motives [317]. Like the previous two categorizations, they include categories for satirical news and fabricated content. They also include propaganda, similar to Tandoc et al. (2017), but explicitly note that it is a type of fabricated content with a political agenda. Additionally, they define *conspiracy theories*, *hoaxes*, and *rumors* as types of misinformation. While all three have distinct technical definitions, these definitions overlap. Conspiracy theories involve unsubstantiated rumors of a conspiracy, making them a type of rumor. Hoaxes typically consist of stories that contain either fabricated content or half-truths. They define *clickbait* as "the deliberate use of misleading headlines and thumbnails of content on the Web" similar to yellow journalism [317]. This clickbait category overlaps with both Tandoc's advertising category [273] and Wardle's misleading content or false connection categories [305]. Their final category is for *biased* news, which is one-sided or hyperpartisan [317]. This also qualifies as a type of misleading content.

Finally, Wang et al. (2020) build upon Tandoc et al.'s typology by adding six additional categories: *clickbait*, alarmist talk (categorized in Table 1.1 as *other*), subjective assumption (categorized as *biased or one-sided*), user-generated news impersonating real news (categorized as *imposter content*), hearsay (categorized as *rumors*), and incorrect content (categorized as *error*) [304].

Many of the categorizations in the literature incorporate the purpose and news context of misinformation, which relate to the misinformation agent's motives rather than just the content's features. For a more precise analysis, I propose separating the purpose and context from the type of misinformation, primarily favoring Wardle et al. (2017)'s categorization [305].

1.3.2 Misinformation Typology

Wardle et al. (2017) define three elements of information disorder: the misinformation **agent** who creates and/or disseminates the misinformation, the type of the **message** and its content, and finally, the **audience** perceptions of the message and the actions they take, if any, as a result of it [305]. These three main elements will organize my proposed misinformation typology.

1. Agent

Various types of actors create or spread misinformation, whether intentionally or not. The agents who initially create misinformation are not always the same as those who spread it, and these actors may have different motivations. According to the literature, there are three main dimensions of actor types: official versus unofficial actors, level of organizational structure, and the use of automated technology [181, 305, 317, 320].

Actor Types: The different types of misinformation creators or disseminators.

- *Official Organizations* - Official actors, including political parties, governments, corporations, and news organizations.
- *Unofficial Organizations* - Unofficial yet well-organized actors could include lobbying groups, terrorist or criminal networks, state-sponsored trolls, or citizen groups.
- *Individuals* - Includes journalists, influencers, or regular people.
- *Bots* - Agents that use automated technology generally can create and disseminate misinformation for cheaper and faster.

Actors have their own purposes and motivations, which are individually not mutually exclusive. For example, one could promote misinformation for both a political agenda and financial reasons. I propose the following primary purposes related to the creation and spread of misinformation, adapted from Wardle et al. (2017) [305] and Zannettou et al. (2019) [317].

Purpose: The purpose of a misinformation message lies in its intention and primary goal.

- *Political* - Misinformation intended to influence political attitudes and opinions. This type of misinformation can include intentionally increasing polarization, inciting violence, and feeding into extremism. This category addresses the *propaganda* and *biased* misinformation types proposed by many of the reviewed papers.
- *Financial* - Misinformation intended to generate clicks and increase revenue, addressing the *advertising* and *clickbait* categories.
- *Distracting* - A message meant to distract the public with a different story, confuse, sow discord, or cause panic. Often, these messages are conspiratorial and intended to promote conspiracy theories, hoaxes, and rumors.
- *Accidental* - False information or context that spreads with no malicious intent. This spreading can occur in a variety of ways:
  - Individuals share something they believe others will correctly determine is a joke rather than something serious [305].
  - Individuals share something they truly believe is accurate [317].
- *Other* - Misinformation created or spread for another reason. This includes sharing misinformation for fun or attention, or attempting to debunk it. Other motivations may involve connecting with others in one's group [305, 317].

2. Message

The stylistic characteristics and news content in misinformation messages created by these actors can vary. I first consider the physical content associated with a piece of misinformation, which includes any text, images, videos, or audio related to the message, such as the headline and body of the article [320]. Actors can use a variety of techniques to generate content or misrepresent it. Some types of misinformation may be easier to counter than others. For example, satirical news sources often disclose that they are satirical, and completely fabricated content or errors can be directly fact-checked. However, misleading or manipulated content can be more insidious. I propose the following categorization of the types of misinformation messages.

Misinformation Types: The type refers to how the message presents misinformation.

1. *Satire and Parody* - Humorous content that typically does not intend to cause harm, though it can fool some (e.g., The Daily Show, The Onion).
2. *False Connection* - Content containing headlines, captions, or images not supported by the rest of the content (e.g., clickbait).
3. *Misleading Content* - Misleading information or opinions presented as facts (e.g., cherry-picking, hyper-partisan news).
4. *False Context* - Correct information shared with false context (e.g., real images with incorrect captions or dates).
5. *Imposter Content* - Information posted while impersonating a genuine source or brand to gain credibility (e.g., the unofficial usage of an official logo or reputable individual's name).
6. *Manipulated Content* - Text, image, or video distortion (e.g., Photoshop, AI-generated images).
7. *Fabricated Content* - A false story, completely made-up.
8. *Error (False Content)* - Generally a mistake that is later corrected, often by a reputable news organization.

In addition to the characteristics that describe how the content is misleading, another important aspect of misinformation messages is their actual news content [320]. The content may be related to the difficulty of debunking misinformation, with some previous studies showing that political misinformation is among the most difficult to debunk [300].

News Topics: The topic is the main item discussed in the message. Five general news categories and a well-known example of misinformation in that domain are listed below.

- *Politics* - Articles on political figures, government policies, elections, and other newsworthy events. Example: "Pope endorses Trump in 2016 election" [256].
- *Health* - Articles related to health and wellness. Topics include pandemics, illnesses, vaccines, and smoking. Example: "Garlic cures COVID-19" [268].
- *Science and Technology* - Articles related to science and research. Topics include climate change, space, and evolution. Example: "The Earth is flat" [5].
- *Business and Finance* - Articles related to businesses, the economy, or consumer products. Example: "Corona beer sales in the U.S. plummet during the COVID-19 pandemic" [67].
- *Entertainment* - Articles related to the entertainment or sports industries, including celebrities, movies, music, and athletes. Example: "Avril Lavigne died in 2003 and was replaced by a clone" [206].
- *Other* - Any other news topics.

3. Audience

Misinformation agents may target either a general or specific audience to influence. The targeted users perceive the misinformation, which they may or may not believe or act upon.

I first consider the target of the misinformation. Audiences are often targeted based on various demographic characteristics, including nationality, age, gender, race, sexuality, religion, income, or country of origin. People may also be targeted based on their membership in a group, such as a consumer group, non-profit, or company [305]. For example, politically driven misinformation may target groups of voters, citizens, or elected officials in a specific state. Other features may also affect the target audience, such as the topics of interest to a user or the size and composition of their friend or follower network [212]. Finally, some misinformation may not be specifically targeted.

Target: The intended recipient of the misinformation.

1. *Demographics* - Specific groups of people based on demographic characteristics.
2. *Group membership* - Specific groups of people based on membership in an organization, such as a political party or company.
3. *Other attributes* - Other targeted features could include individuals interested in specific topics or highly influential users.
4. *General public* - Untargeted misinformation.

Next, I consider how the audience receives the message. Individuals determine whether to accept a message, either entirely or partially, based on their evaluations of the content and supplier. For example, is the supplier a trusted messenger for the targeted individual? Is the language in the message highly emotional? According to previous research, individuals take multiple factors into account when evaluating content, including familiarity, simplicity, perceived credibility, motivated reasoning, and the persuasiveness of the message [181, 305]. Individuals have varying levels of misinformation susceptibility, often depending on these factors, which will be explored in later chapters.

Reception: Regardless of the reasons behind it, seminal work on reception theory posits that individuals perceive messages in one of three ways: [124, 305].

1. *Full Acceptance* - Accept the message as is.
2. *Partial Acceptance* - Accept parts of the message.
3. *Rejection* - Reject the message in its entirety.

Finally, I consider whether the targeted group or individual takes an action, specifically on social media. An individual could choose to ignore the message. They could like or comment positively if they agree with the message in part or in full, boosting the post's engagement metrics. They could additionally reshare the message to a larger audience, furthering its spread. If an individual disagrees with the message, they could report or flag the post, comment critically, or reshare the message in a critical manner. These actions are further investigated in Chapters 3 and 4.

Actions: The actions an individual can take on social media networks in response to being exposed to misinformation.

- *No Action* - Ignore the message
- *Positive Action* - Engage with the message positively, such as by liking, commenting, or resharing.
- *Negative Action* - Engage with the message negatively, such as by commenting, adding context, or reporting the message.

In summary, there are three elements to misinformation: the agent, the message, and the audience. Table 1.2 summarizes the various aspects described in this categorization.

Table 1.2: Proposed misinformation typology.

Element	**Features**	**Examples**
Agent	Actor Types	Official or unofficial organizations, individuals, bots
	Purpose	Political, financial, distracting, accidental, other
Message	Misinfo Type	Satire/parody, false connection, misleading content, false context, imposter content, manipulated content, fabricated content, errors
	News Topic	Politics, health, science and technology, business and finance, entertainment, other
Audience	Target	Demographics, group membership, attributes, general public
	Reception	Full acceptance, partial acceptance, rejection
	Actions	No action, positive action, negative action

1.3.3 Misinformation Pipeline

In addition to understanding the three main elements of misinformation, it is important to also consider the lifecycle of a specific piece of misinformation content on social media, often referred to as the “misinformation pipeline” [72]. This section explores how the proposed misinformation typology informs the misinformation pipeline, and how we can analyze the ways in which misinformation is created and spread on social media in order to develop targeted interventions and countermeasures.

There is no consensus in the academic literature on the best way to define the social media misinformation pipeline. According to Wardle et al. (2017), there are three phases: the original *creation* of the message, the *production* (and reproduction) of the message into shareable content, and the *distribution* of the message to the public or intended audience [305]. Similarly, Ciampaglia (2018), in their comprehensive discussion article on a proposed research agenda for the academic community, defines the pipeline as occurring in three stages: *production* or creation of the misinformation content, the *diffusion* or distribution of the content, and the *verification* of the content, which refers to the actions taken after the misinformation has spread [72]. Finally, Ng and Taeihagh (2021) define a four-stage framework that describes how misinformation can be spread intentionally on social media via the usage of APIs: *network creation*, such as by creating an account and following others, *profiling* a target audience, *content generation* and the production of the misinformation, and *information dissemination*, which refers to actions taken to effectively spread the media content [212].

These previous definitions all include the creation or production of a message and the spread of the message, while some also address what happens after the message is spread. I propose combining these ideas and defining the following pipeline of the misinformation lifecycle into three main phases: **creation**, **spread**, and **belief**.

1. Creation

Creation typically refers to the original inception of the misinformation message. Another important aspect to consider is not just the generation of the content but also the establishment of accounts, or multiple accounts, that will disseminate it, along with the networks they form to further spread the message, as noted by Ng and Taeihagh in their framework [212]. I define creation as having two related components: network creation and content creation.

- *Network Creation* - This component refers to the creation of accounts and networks that will later be used to spread misinformation [212]. Users typically need to create an account on a social media platform to post content on that platform. Users sometimes use their real name or a screen name, and they may disclose their location, profession, interests, or describe a specific purpose behind their account [155, 322]. After account creation, users can begin following and friending other accounts on the platform, thus creating their own social network, as well as liking posts or topics of interest. Malicious users may obtain or hijack existing accounts, or create a set of coordinating bot accounts, and start developing the networks of these accounts such that they can later spread the misinformation they generate [212].
- *Content Creation* - Content creation is the process of developing the original misinformation message and transforming it into a media product [305]. This includes both text generation and the creation or curation of related images, audio, or video content [212]. Generative AI can assist in the rapid development of content and in constructing targeted messages that resonate with specific audiences [212].

Potential interventions in the creation step of the misinformation pipeline typically focus on algorithmic detection of the misinformation content [72], or the detection of fake accounts and inauthentic activity [212]. For example, platforms typically seek to block spam, bot, or fake accounts, and they use CAPTCHAS or other methods, such as requiring a phone number, to prevent the easy creation of fake accounts [212].

2. Spread

Spread refers to how users, either intentionally or unintentionally, distribute or disseminate the misinformation content [72, 305]. This phase begins when one or more accounts post the misinformation content, and it can be further amplified to additional audiences in numerous ways. I define spread as having two related components: sharing and amplification.

- *Sharing* - This component refers to the direct act of sharing the misinformation content with others, such as by posting the content, messaging the content directly to specific users, or forwarding the message.
- *Amplification* - Amplification refers to how engagement with the content or algorithmic bias can further spread the message [72]. Users can comment on their own posts, reply to direct messages, or add specific keywords or hashtags to further its spread. Malicious users can also use coordinated efforts with other users or bot accounts to artificially increase engagement with a post to further its initial spread [212].

Potential interventions in this step of the misinformation pipeline typically focus on introducing friction before regular users share content without thinking [167] or implementing algorithmic downranking and other platform alterations to reduce artificial amplification [305].

3. Belief

Belief refers to the false beliefs that may arise from the spread of this misinformation content. Once the misinformation has spread widely, actions can be taken to address the false content [72]. However, prevention is another important aspect to consider when analyzing beliefs [180]. Prevention can occur at any point in the pipeline or external to it through educational efforts, such that even if the misinformation is widely spread and viewed, users are already aware that they should not believe it. I define belief as having two related components: verification and prevention.

- *Verification* - This component refers to the verification of the posted and spread content. Verification can happen via the usage of automated or human fact-checkers [72]. Human fact-checking can involve journalists, experts, moderators, or other users. Fact-checking can result in content labels, warnings, or removal.
- *Prevention* - Prevention includes inoculating or warning users about specific misinformation content or techniques they may encounter [180] as well as any educational effort designed to enhance literacy skills and overall competencies [141]. There are various aspects of literacy, including general media literacy (the ability to critically assess and evaluate a piece of media), information literacy (the ability to search for and find relevant information), news literacy (the ability to identify and analyze the news), and digital literacy (the ability to find and understand online information) [146].

Potential interventions in this step of the misinformation pipeline typically focus on correcting, retracting, or removing false content [300], or educating and warning the public about the misinformation they may encounter [180].

In summary, there are three phases in the misinformation pipeline on social media for a piece of misinformation: creation, spread, and belief. Table 1.3 summarizes the various components of the pipeline.

Table 1.3: Proposed misinformation pipeline.

Phase	Component	Description
Creation	Network	Create accounts, friend/follow others
	Content	Write message/post, create images/audio/video
Spread	Sharing	Post content, share content with other users
	Amplification	Comment/reply to messages, add keywords/locations/hashtags
Belief	Verification	Automated or human fact-checking, labeling
	Prevention	Inoculation, warnings, literacy (news, digital, media, information)

1.4 Countermeasures Categorization

This section first presents an overview and comparison of how researchers classify misinformation interventions. Then, building upon this prior work, I outline a proposed categorization.

1.4.1 Literature Review

Several review papers and meta-analyses have been published in the field of misinformation interventions. Some reviews, such as Helmus and Keppe (2021) from the Rand Corporation, focus solely on policy papers [131]. Most others examine specific intervention categories, such as content moderation [143] or media literacy [141], rather than multiple categories.

Four of the most comprehensive review articles are analyzed: Courchesne et al. (2021) [79], Aghajari et al. (2023) [8], Blair et al. (2024) [46], and Kozyreva et al. (2024) [167]. These articles were selected for their recency, the breadth of interventions covered, and their diverse disciplines. The Courchesne et al. article discusses platform interventions and was published in the Harvard Misinformation Review, an interdisciplinary journal mainly associated with social sciences, as defined by the SCImago journal rank database[1]. The article by Aghajari et al. was published in a computer science conference proceedings. The Blair et al. article highlights research from both the Global North and the Global South, and it was published in a psychology journal. Finally, the Kozyreva et al. article was recently published with numerous high-profile scholars in this research area and appeared in the prestigious Nature Human Behavior journal. Table 1.4 summarizes the interventions analyzed by each paper in alphabetical order.

The Courchesne et al. article classifies types of interventions based on those publicly announced as having been implemented by various social media platforms [79]. These 10 categories are advertisement policy, content labeling, content / account moderation, content reporting, content distribution / sharing, disinformation disclosure, disinformation literacy, redirection, security / verification, and other. The article notes that fact-checking was included in the disinformation disclosure category, so this is categorized as fact-checking in Table 1.4.

The Aghajari et al. article categorizes interventions primarily based on the driver of the misinformation that each intervention targets: Content, Source, Individual Users, or Community [8]. For content-based interventions, they discuss fact-checking, warning labels, and platform

[1] https://www.scimagojr.com

alterations (reducing the size of misinformation). For source-based interventions, they discuss source credibility labels and the crowdsourcing of those labels. For user-based strategies, they consider media literacy, accuracy prompts, and account removal. Lastly, for community-based interventions, they address social norms.

Table 1.4: Categorizations of misinformation interventions in the literature.

Type	**Courchesne et al. (2021) [79]**	**Aghajari et al. (2023) [8]**	**Blair et al. (2024) [46]**	**Kozyreva et al. (2024) [167]**
Account Moderation	×			
Account Removal		×		
Accuracy Prompts		×	×	×
Advertising Policy	×			
Content Distribution	×			
Content Labeling	×			
Content Moderation	×			
Context Labels			×	
Crowdsourcing		×		
Debunking			×	×
Fact-Checking	×	×		×
Friction			×	×
Inoculation			×	×
Journalist Training			×	
Lateral Reading				×
Media Literacy	×	×	×	×
Platform Alterations		×	×	
Politician Messaging			×	
Redirection	×			
Reporting	×			
Security/Verification	×			
Social Norms		×	×	×
Source Credibility Labels		×	×	×
Warning Labels		×		×
Other	×			

The Blair and Kozyreva articles [46, 167] categorize interventions not just in terms of which parts of the platforms are affected, but also what part of the social media misinformation pipeline is targeted (such as the creation, spread, or belief in misinformation - see Table 1.3). More specifically, the Blair et al. article categorizes 11 types of interventions into four main groups:

- *Informational* - For informational interventions, they define inoculation / prebunking, debunking, credibility labels / tags, and contextual labels / tags. These interventions aim to correct misinformation. The part of the misinformation pipeline targeted is (**belief - verification**).

- *Educational* - Under educational interventions, they define media literacy. This category of interventions seeks to prevent belief in misinformation (**belief - prevention**).
- *Sociopsychological* - In this category, they define accuracy prompts, friction, and social norms. These interventions aim to discourage users from spreading misinformation (**spread - sharing**).
- *Institutional* - Lastly, they define platform alterations, politician messaging, and journalist training under institutional interventions. This fourth category of interventions targets the behavior of the original creators and distributors of misinformation (**creation - content, spread - amplification**).

The Kozyreva et al. article classifies nine types of interventions into three main categories [167]:

- *Nudges* - Under nudges, they include accuracy prompts, friction, and social norms. These interventions address sharing behavior by discouraging users from further distributing misinformation (**spread - sharing**).
- *Boosts and Educational Interventions* - This category includes inoculation, lateral reading and verification strategies, and media literacy tips. These interventions aim to increase user competence in evaluating content online and discerning misinformation or untrustworthy content from reputable information (**belief - prevention**).
- *Refutation Strategies* - Lastly, the refutation strategies category includes debunking and rebuttals, warning and fact-checking labels, and source credibility labels. This group targets belief in misinformation (**belief - verification**).

Many of the review articles used a similar categorization of countermeasures; however, there is no common typology [8, 46, 79, 167]. Many of these defined categories overlap (e.g., lateral reading skills vs. media literacy) or are sub-categories of other categories. For example, redirection is a form of content distribution. Similarly, context labels and source credibility labels are types of content labeling.

Additionally, specific categories are absent, particularly user-led or institutional interventions such as government regulation. User-based measures are an often overlooked aspect in the fight against misinformation. Individual-level debunking, especially from trusted messengers, is effective in various contexts [28, 48, 182, 275, 303]. In addition to user reporting and social norms, users can block others or engage in social corrections. When misinformation is successfully posted on social media, other users serve as the first line of defense since they can flag or debunk it. While social corrections can be considered a type of debunking or fact-checking, the social context may be important to distinguish. User-based countermeasures are addressed in greater detail in Chapters 3 and 4.

1.4.2 Countermeasures Typology

After reviewing the literature and these previous categorizations, I developed eight general categories of countermeasures, as shown in Table 1.5, primarily classified by the part of the platform affected. The first five general categories mainly focus on platform-driven interventions, although in some instances, governments exercise oversight through regulation or other legislative measures. Platform interventions often involve algorithmic changes regarding what content can be created or distributed on the platform (content distribution, content moderation, account moderation) or front-end design changes concerning how content is displayed or can be interacted with after it has spread (content labeling, user-based measures). Media literacy and other institutional measures typically require more upfront investment and can be implemented by platforms, governments, or various civic institutions.

Table 1.5: Proposed categorization of misinformation interventions.

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

The proposed typology introduced in this section defines the general categories of misinformation interventions. This typology is helpful in analyzing the literature in this field and will be used in the topic and citation network analysis conducted in Chapter 2. See Appendix A for more details on each intervention category, specific intervention types, and associated references.

1.4.3 Targeting the Misinformation Pipeline

This section discusses how the countermeasures categorization informs the development of targeted interventions for each part of the misinformation pipeline.

Blair et al. (2024) define four categories of interventions that each target a different aspect: original creation and distribution (*institutional*), additional distribution by regular users (*sociopscyhological*), correction of false beliefs (*informational*), and prevention of belief in misinformation (*educational*) [46]. Kozyreva et al. (2024) likewise focus on three parts of the misinformation pipeline: discouraging spread by users (*nudges*), correcting false beliefs (*refutation strategies*), and preventing belief in misinformation by increasing competencies (*boosts and educational interventions*) [167].

Table 1.6 summarizes how interventions in each of the eight categories target and combat misinformation in the misinformation pipeline. When considering the three primary phases of the pipeline, there are several intervention points available. More specifically, an intervention can target the **creation**, **spread**, or **belief** in misinformation. Belief can be addressed after the fact, such as by debunking or correcting misconceptions, or proactively through prevention efforts and increased competency.

Table 1.6: The categories of countermeasures and the phases of the misinformation pipeline they target.

Phase	Component	Intervention Category
Creation	Network	Account Moderation
	Content	Content Moderation
Spread	Sharing	Content Distribution
	Amplification	Content Distribution, Content Moderation, Account Moderation
Belief	Verification	Content Moderation, Content Labeling, User-based Measures
	Prevention	Media Literacy and Education, Institutional Measures

The **creation** phase is typically targeted by account and content moderation techniques, such as restricting or limiting users or specific types of content. The **spread** phase is targeted primarily through content distribution techniques, although automated content moderation can also help reduce algorithmic amplification. Finally, countermeasures can target **belief** in posted misinformation by promoting user-based measures like social corrections or reporting features, while also focusing on prevention through the promotion of social norms, media literacy, educational efforts, and external institutional measures.

1.5 Platform Policies

The leading social media platforms currently define and address misinformation in different ways, varying in the types of content they prohibit or limit and how they counter it. This section reviews current social media policies and compares the types of misinformation that are disallowed, along with the interventions typically used.

1.5.1 Community Guidelines

Before comparing the ways platforms respond to misinformation, I first investigate their community guidelines to determine what constitutes prohibited content. More specifically, I review the policies of the most popular platforms in the U.S. as of 2024 based on the percentage of U.S. adults who report using each platform. The top five are YouTube (85%), Facebook (70%), Instagram (50%), Pinterest (36%), and TikTok (33%) [263]. Table 1.7 summarizes the misinformation policies of the top five platforms, categorized by the specific elements of misinformation that they address (as defined in Table 1.2).

All platforms target false information from agents by banning inauthentic accounts (*actor types*) that engage in deceptive practices or scams (*financial* purpose)[2] [3] [4]. For example, YouTube

[2] https://transparency.meta.com/policies [Accessed 03-21-2025

[3] https://policy.pinterest.com/en/community-guidelines [Accessed 03-21-2025]

[4] https://www.tiktok.com/community-guidelines/en [Accessed 03-21-2025]

prohibits impersonation, fake engagement, and spam[5]. Additionally, dangerous individuals or organizations, such as criminal groups, violent extremists, or terrorist supporters, are generally banned, not only their illegal content.

Table 1.7: Current platform policies regarding misinformation.

Element	Features	Policy	YouTube	Meta	Pinterest	TikTok
Agent	Actor Types	Fake accounts	×	×	×	×
		Dangerous accounts	×	×	×	×
	Purpose	Spam	×	×	×	×
Message	Misinfo Type	AI Disclosure	×	×		×
	News Topic	Health	×	×	×	×
		Elections	×	×	×	×
		Science			×	×
Audience	Target	Hate speech	×	×	×	×

Platforms vary in how much offline harm false or misleading content can cause before determining whether to take action on specific messages. Most platforms generally prioritize the news topic of the message rather than its type, and the majority clearly indicate that misinformation posing a serious risk of direct offline harm is prohibited. However, the two most prominent platforms, YouTube[6] and Meta[7], direct their misinformation policies primarily on two news topics. Specifically, they focus on **health misinformation** (such as fake cures and misleading claims about vaccines and public health) and **election or civic misinformation** (such as false election dates, suppression of voter groups, or any efforts intended to undermine electoral integrity or census participation).

Other platforms, such as Pinterest[3] and TikTok[8], take a broader approach. TikTok states it prohibits false or misleading content that "may cause serious harm to individuals or society, regardless of intent." Pinterest indicates that it restricts or moderates content that may harm someone's "well-being, safety or trust." Acknowledging the erosion of trust and potential societal impacts as forms of harm is significant, as it reflects a more comprehensive strategy than most platforms. Additionally, both Pinterest and TikTok moderate election and health misinformation like other platforms, but they also include **science misinformation**, such as climate change denial, as prohibited content. Lastly, both platforms disallow harmful conspiracy theories.

YouTube[9], Meta[7], and TikTok[8] also require that AI-manipulated content that is realistic in nature or has the potential to mislead is labeled as such. Pinterest currently lacks an explicit

[5] https://support.google.com/youtube/answer/2801973?hl [Accessed 03-21-2025]

[6] https://www.youtube.com/howyoutubeworks/our-commitments/fighting-misinformation/ [Accessed 03-21-2025]

[7] https://transparency.meta.com/policies/community-standards/misinformation [Accessed 03-21-2025]

[8] https://www.tiktok.com/community-guidelines/en/integrity-authenticity [Accessed 04-23-2025]

[9] https://support.google.com/youtube/answer/14328491?hl [Accessed 03-21-2025]

AI disclosure policy as of March 2025; however, it bans fabricated or AI-generated content that erodes trust or causes harm[3]. Finally, all platforms ban hate speech and harassment, which are often aimed at specific audiences and demographic groups, including protected groups.

1.5.2 Enforcement

In addition to differing community guidelines surrounding misinformation, platforms implement a range of interventions, engaging in all the intervention categories described in Table 1.5. For example, YouTube[6] outlines four primary strategies they use for combatting misinformation:

- **Reducing the spread** of potentially harmful content through interventions like algorithmic downranking and automated misinformation detection (*Content Distribution*, *Content Moderation*).
- **Increasing the distribution of high-quality content** by elevating authoritative sources and using fact-checking, warning labels, and source credibility labels (*Content Labeling*).
- **Rewarding trusted accounts** with monetization policies and demonetizing those who violate community guidelines (*Account Moderation*).
- **Giving users control** through reporting tools, blocking and filtering abilities, media literacy initiatives, and institutional measures like media support (*User-based Measures*, *Media Literacy and Education*, *Institutional Measures*).

According to its Community Standards[7], Meta also employs misinformation detection and content distribution strategies to slow the spread of hoaxes, provides resources to improve media literacy, and requires disclosure of AI-generated content. However, it recently removed fact-checking in the U.S. and replaced it with context labels [149]. TikTok implements algorithmic content moderation, uses professional fact-checkers, labels government accounts, and requires that realistic AI-generated content is labeled[8]. Meanwhile, Pinterest focuses on removing or limiting the distribution of violating content and regularly releases transparency reports[3].

Reviewing platform policies validates the proposed typologies of misinformation and countermeasures and helps us narrow our scope. Given that the platforms primarily focus on health and political misinformation as the most potentially harmful forms worth addressing, this dissertation will concentrate primarily on these two types of misinformation.

1.6 Data

This thesis creates and uses five primary datasets to characterize countermeasures while considering several relevant factors. All datasets are summarized in Table 1.8.

Literature Corpus

The *Literature Corpus* data is a collection of papers pulled from four prominent review papers and additional keyword searches to analyze gaps in the literature. The topic and citation network

Table 1.8: Summary of the datasets used in this thesis.

Data	**Size**	**Dates Collected**	**Ch 2**	**Ch 3**	**Ch 4**	**Ch 5**	**Ch 6**
Literature Corpus	451 papers	2004-2025	✓				✓
Countermeasures Survey	1010 responses	12 Jul to 23 Aug 2024		✓		✓	✓
Training Quizzes	23 responses	5 Feb 2024			✓		
Training Posts	40 posts	Various			✓		
Expert Opinions Survey	39 responses	23 Jan to 24 Mar 2025					✓

analysis of this dataset is conducted in Chapter 2. Additionally, this data is summarized alongside the other datasets in Chapter 6.

Countermeasures Survey Data

The *Countermeasures Survey* data was collected during the summer of 2024. Behavioral and opinion-based questions were asked of 1010 American-based social media users. This data set includes standard demographic questions, as well as questions about behaviors and opinions related to user-based, platform-based, and government-level countermeasures. The questions related to user-based interventions are analyzed in Chapter 3, while those related to platform and government interventions are analyzed in Chapter 5. Additionally, this data is summarized alongside the other data sets in Chapter 6.

Pre and Post-Training Quizzes

The *Training Quiz* data includes responses from both a pre-training and post-training quiz collected from 23 government analysts. The respondents completed a pre-quiz to assess their knowledge of misinformation detection and countering ability. They then underwent relevant training. Finally, they took a post-training quiz to see if there was any improvement in their detection abilities. This data set is used in Chapter 4.

Social Media Training Posts

To create the training quizzes in Chapter 4, we curated a set of *Social Media Training Posts*. These posts include a variety of misinformation, conspiracy theories, pink slime, and accurate news items to assess participants' misinformation detection and countering skills. We used fact-checking websites, such as FactCheck.org, to identify relevant misinformation posts or searched social media platforms directly for content on specific topics. We also reviewed existing COVID-19 Twitter datasets previously collected by the CASOS research group. Most of these posts came from X/Twitter and Facebook.

Expert Opinions Survey

The *Expert Opinions* data set is a survey of misinformation researchers designed to gather their opinions on all operationalized countermeasures defined and analyzed in this thesis, along with the characteristics of those countermeasures, including effectiveness, acceptance, effort level, political feasibility, and cost. This survey was designed for analysis in Chapter 6.

1.7 Chapter Overview

This thesis establishes a framework for assessing and developing practical and effective countermeasures to misinformation. Chapter 2 conducts a bibliometric analysis to better understand the current state of the literature and identify research gaps. User-based misinformation interventions are explored in Chapters 3 and 4. Chapter 3 outlines current user opinions and behaviors regarding this subject and considers how this information could be leveraged to improve user-based countermeasures. Chapter 4 builds directly on Chapter 3, focusing specifically on improving media literacy interventions, a type of user-based intervention, to determine whether they can be adapted to improve countering abilities (not just misinformation detection abilities) among a highly skilled audience. Chapter 5 discusses platform and government interventions as well as the factors most associated with public support for these interventions: perceived fairness, effectiveness, and intrusiveness. Chapter 6 synthesizes information from the previous chapters and a survey conducted with misinformation experts to provide analysis-driven recommendations. Finally, I conclude with a discussion of the results, recommendations, limitations, and potential avenues for future work in Chapter 7.

Chapter 2

Bibliometric Analysis of the Interventions Literature

To supplement the literature review in the Introduction, I conducted a systematic scoping review and a bibliometric analysis of relevant papers to generate a more comprehensive contextual background on the current literature in this field.

In this chapter, I start with four prominent review papers published in different research disciplines and compile all relevant papers they reviewed, along with newly published papers, leading to a total of 451 articles. I analyzed this collection of publications using a variety of metrics to provide policymakers and academics with relevant information on the current state of the literature in this domain, to identify any research gaps, and to focus the remaining chapters of this dissertation.

The main research question for this chapter is: How are interventions to combat misinformation currently studied in the academic literature? More specifically:

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. What types or categories of interventions have been studied the most, and how has this changed over time?
3. What set of impacts has been researched? The primary impacts of an intervention include its effectiveness and level of user acceptance.

2.1 Introduction

Most of the literature in this field, including the review articles discussed in the Introduction, focuses on testing countermeasures and analyzing their effectiveness in reducing the creation, spread, or belief in misinformation. However, this focus overlooks the equally important metrics of practicality and acceptability to users. The review articles led by Courchesne and Blair explicitly included only experimental studies and excluded nonexperimental ones [46, 79]. Similarly, the Kozyreva article included only empirical studies, providing evidence of the efficacy of the interventions studied [167]. Only the review article led by Aghajari included papers that simply presented an intervention rather than requiring studies to examine their effectiveness [8].

Although the four selected review articles were among the most comprehensive in the literature, their emphasis on effectiveness likely restricted the types of interventions covered to primarily platform-based rather than institutional or user-led interventions. I aim to address this gap by considering user-level, platform-level, and policy interventions. This review starts with these four prominent review articles as seed papers and then supplements them with additional research from underrepresented intervention areas. The objective of this review is to gain insights into the misinformation intervention landscape through a bibliometric analysis of relevant articles.

2.2 Data and Methods

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 answer these research questions. We followed the methods used in two previous computer science review articles focusing on interventions [8, 316]. More specifically, we adhered to the modified PRISMA guidelines for scoping reviews [283], as well as the more specific guidelines for systematic reviews in the information systems field developed by Okoli [217].

Given our research questions and objectives, we developed a protocol for this review that outlined in advance the steps and the procedures we would carry out. We developed the intervention labels and eligibility criteria and defined our literature search strategy. Next, we trained the labelers, labeled the papers, and synthesized the results.

2.2.1 Paper Labels

We developed a comprehensive list of 35 specific interventions based on the general categories defined earlier in Table 1.5. Refer to Table 2.1 and Appendix A for more detailed definitions and references for each specific intervention. Note that each general category was always included as a specific intervention label. This label is applied when a paper studies an intervention that falls within that general category but is not explicitly defined by the other specific intervention categories. In addition to the specific intervention labels, we defined four additional 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 in 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.

Each paper was labeled by the countermeasures it discussed or analyzed. Papers were also assigned the *review article*, *meta-analysis*, *effectiveness*, and *acceptance* labels where appropriate. It is important to note that these labels are not mutually exclusive, as some papers can cover

Table 2.1: Intervention topic labels.

General Category	**Specific Intervention**	**Definition**
Content Distribution	Content Distribution	The distribution of content on social media
	Redirection	Redirecting users to other content when searching
	Accuracy Prompts	Reminding people about accuracy
	Friction	Pause and reflect before engaging with content
	Platform Alterations	Altering how content is distributed or displayed
	Advertising policy	What ads are shown to which users
Content Moderation	Content Moderation	How content is shown or removed on social media
	Fact-Checking	Verification of information, often by experts
	Debunking	Fact-checking with context, narrative coherence
	Algo. Content Moderation	Algorithmic content moderation, like downranking
	Misinformation Detection	Automated detection of misinformation
Account Moderation	Account Moderation	Moderating through suspensions, bans, demonetization
	Account Removal	The removal of a user from one or more platforms
	Shadow Banning	Limiting the spread of posts from certain accounts
Content Labeling	Content Labeling	A type of misinformation disclosure through labels
	Crowdsourcing	Using regular people to verify and label information
	Warning Labels	General warnings about misinformation
	Source Credibility Labels	Disclosing or labeling a post's source or credibility
	Context Labels	Adding context to a post, like via community notes
User-based Measures	User-based Measures	How people respond to seeing misinformation
	Reporting	Users can report users or their posts
	Blocking	Users can block users or specific topics
	Social Corrections	Users that fact-check/debunk other users
	Social Norms	Using peer or community influence to change behavior
	Retractions	When accounts retract misinformation they posted
Media Literacy and Education	Media Literacy	Improve the public's civic or digital reasoning
	Fake News Games	Games designed to help people detect misinformation
	Inoculation	Prebunking misinformation before exposure
Institutional Measures	Institutional Measures	Measures by civic society, governments, or institutions
	Media Support	Investing in local news, journalist training
	Data Sharing	Sharing high-quality data with researchers
	Government Regulation	Any relevant laws, rules, or regulations
Other	Other	New interventions or those not fitting any category
	Generative AI	The usage of gen AI to counter or detect misinfo
	Combining Interventions	Using multiple interventions at once

multiple interventions. Furthermore, although all interventions are assigned to one general category, some may apply to multiple categories. For example, fact-checking is a type of content moderation that is sometimes implemented as content labels.

2.2.2 Inclusion Criteria

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

1. **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.
2. **Article Type:** The article is a research article. It is not an opinion piece, research proposal, or simply an abstract.
3. **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:
   - *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.

2.2.3 Literature Search

Our literature search was conducted in two stages. Figure 2.1 outlines our literature selection and review process. First, we gathered all the studies analyzed by our four "seed papers." We removed duplicate papers, applied the exclusion criteria, and labeled the papers according to the definitions provided in Section 2.2.1 and Table 2.1. This process resulted in 365 labeled papers by the end of Stage 1.

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 in 2024 or 2025

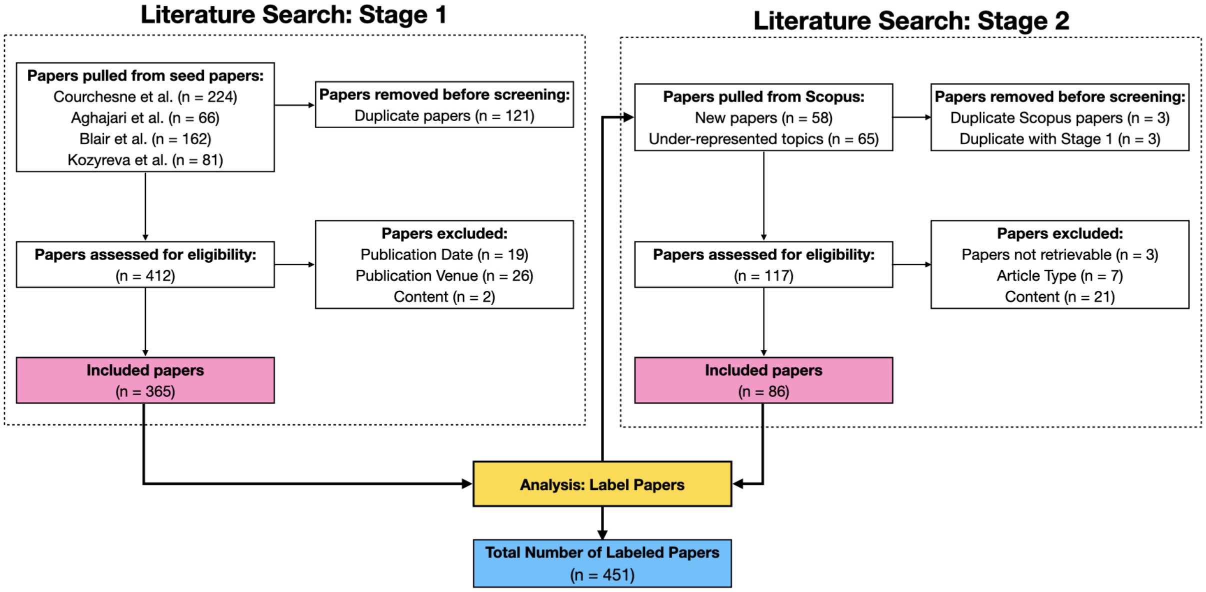

Literature Search: Stage 1
Papers pulled from seed papers:
Courchesne et al. (n = 224)
Aghajari et al. (n = 66)
Blair et al. (n = 162)
Kozyreva et al. (n = 81)
Papers removed before screening:
Duplicate papers (n = 121)
Papers assessed for eligibility:
(n = 412)
Papers excluded:
Publication Date (n = 19)
Publication Venue (n = 26)
Content (n = 2)
Included papers
(n = 365)
Literature Search: Stage 2
Papers pulled from Scopus:
New papers (n = 58)
Under-represented topics (n = 65)
Papers removed before screening:
Duplicate Scopus papers (n = 3)
Duplicate with Stage 1 (n = 3)
Papers assessed for eligibility:
(n = 117)
Papers excluded:
Papers not retrievable (n = 3)
Article Type (n = 7)
Content (n = 21)
Included papers
(n = 86)
Analysis: Label Papers
Total Number of Labeled Papers
(n = 451)

Figure 2.1: Literature review process.

through 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 2.2 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.

2.2.4 Training of Labelers

Three advanced high school interns and ChatGPT assisted us with labeling for this project. The interns received individual training, as they began at different times in the summer of 2024. A training PowerPoint was created and shared with the interns for their reference. A summary of the training is provided below:

- **Background and Motivation:** Interns received an overview of misinformation and countermeasures. The literature review in Sections 1.3 and 1.4 was summarized.
- **Project Overview:** We discussed the project goals and research questions, and reviewed a summary of the 8-step systematic literature review guide as defined by Okoli [217].

Table 2.2: The intervention labels used the least after Stage 1, sorted from highest to lowest according to the number of papers assigned to those labels.

Intervention Label	**Scopus Keywords Used**
Context Labels (9)	(“context” AND “label*”) OR (“community” AND “note”)
Algorithmic Content 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”) OR “monetiz*”
Blocking (0)	“blocking”
Data Sharing (0)	“data sharing”
Government Regulation (0)	(“government” AND “regulation”) OR (“government policy”)
Generative AI (0)	“gen* AI” OR “chatbot”

- **Reading Academic Papers:** Interns received training on effectively reading and understanding academic papers using the CIMO framework and the SQ3R method for reading comprehension [217].
  - *CIMO Framework* - The Context-Interventions-Mechanisms-Outcomes framework outlines the primary questions they should consider when reading the papers. This framework considers the context and systems that are involved, the event or intervention being studied, the relationship between the intervention and the outcomes, and finally, the outcomes of the intervention [49].
  - *SQ3R Method* - The Survey, Question, Read, Recall, and Review technique is used to assess reading comprehension. During the survey step, interns were instructed to skim and scan relevant parts of each paper, including the title, abstract, keywords, and, if necessary, the introduction and conclusion, to determine whether the paper should be included. In the questioning step, they were asked to evaluate the paper’s relevance while keeping the CIMO framework in mind. Next, they were taught to read and make connections with previous papers. Finally, they were asked to recall the paper’s main points by reviewing their notes and writing a brief summary of those points [245].
- **Reference Manual:** Interns were given a checklist describing the inclusion and exclusion criteria, along with detailed definitions for all the labels defined in Section 2.2.1.
- **Tabular Documentation:** Interns were provided with a Google Sheets document to summarize the main findings, specify the interventions studied, indicate whether each paper studied effectiveness and acceptance, and determine if each was a meta-analysis or review article, among other items.

2.2.5 ChatGPT Prompts

ChatGPT prompts were designed to assist with labeling the specific interventions studied and whether the papers studied effectiveness or acceptance. These prompts were developed and refined after all human labelers completed their labeling tasks. We applied standard prompt-engineering best practices to create the prompts, including clearly articulating the task, experimenting with different phrasing and techniques, and testing various model versions [7]. Refer to Appendix B for the final prompts.

Twenty Stage 1 papers were chosen as a test set while refining the prompt. These papers were selected to span as many intervention categories as possible. The prompts for labeling each paper's effectiveness and acceptance were executed using ChatGPT's 4o-mini and 4o models. Both models achieved the same accuracy for the effectiveness task, agreeing on 19 of the 20 papers. For the acceptance task, the 4o-mini model agreed with the final label on 17 papers, while the 4o model agreed on 18 papers. Calculating Cohen's kappa inter-rater agreement values would be inappropriate in this context because the categories were unbalanced (all papers studied effectiveness but not acceptance), which would result in all kappa values being 0 [75].

While the 4o model had a slightly higher percent agreement for acceptance, it is significantly more computationally intensive and financially costly. In February 2025, when these models were run via the API, OpenAI charged $0.15 per one million input tokens and $0.60 per one million output tokens for the 4o-mini model. However, they charged $2.50 per one million input tokens and $10 per one million output tokens for the 4o model[1]. A token roughly equals four English characters or three-quarters of one word[2]. Considering we were using ChatGPT to process academic papers that are typically thousands of words long, the 4o model was avoided whenever possible. Therefore, the effectiveness and acceptance prompts were run on the rest of the papers using ChatGPT's 4o-mini model.

For the intervention labeling task, there were 35 potential labels, which created significant room for disagreement. We improved the intervention prompt by making the definitions concise and explicitly stating at the beginning of the prompt that interventions should be labeled only as defined in the prompt. We tested three techniques to improve agreement with the final labels.

1. **Separate Category Prompts** - Creating eight prompts, one for each general category. These results were the worst, regardless of how the prompts were phrased. Using this technique, ChatGPT tended to assign too many labels to the papers. This over-labeling may have occurred because many of the general categories overlap (for example, fact-checking could be considered a content moderation technique or a content labeling technique).
2. **Alphabetical Combined Prompt** - Developing a single prompt and arranging the interventions in alphabetical order. These results were an improvement.
3. **Ordered Grouped Combined Prompt** - Developing a single prompt and ordering the interventions by category, much like how they are organized in Table 2.1. These results were the best.

The intervention labeling was run using both 4o-mini and 4o for each of the three techniques.

[1] `https://platform.openai.com/docs/pricing` [Accessed 03-13-2025]

[2] `https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them` [Accessed 03-13-2025]

Table 2.3: Summary of interventions labeling by different prompt and ChatGPT model types. The values indicate Jaccard similarity scores that compare the labeled output produced by the model run against the final label.

		Separate Prompts		Alphabetized Prompt		Ordered Prompt	
Paper	**Final Labels**	**4o-mini**	**4o**	**4o-mini**	**4o**	**4o-mini**	**4o**
[9]	advertising policy, debunking	0	0.25	1	0.5	1	0.5
[17]	accuracy prompts, crowdsourcing, media literacy	0.176	0.3	1	1	0.667	1
[28]	social corrections, source credibility labels	0.111	0.143	0.333	0.5	0.5	0.5
[29]	friction	0	0	0	0	0	0
[47]	redirection	0	0	0	0.333	0.333	1
[47]	algorithmic content moderation, social corrections	0.111	0.222	0.333	1	0.333	0.333
[95]	retractions, source credibility labels	0.25	0.5	0.25	0.25	0.5	0.333
[98]	combining interventions, debunking, social norms	0.333	0.5	0.25	0.25	0.25	0.25
[101]	accuracy prompts	0.143	0.167	1	0.5	0.5	1
[105]	context labels, source credibility labels	0	0	0	0.333	0	0
[106]	fact-checking, inoculation	0.111	0.4	0.667	0.667	0.667	1
[146]	media literacy	0.25	0.5	0.5	1	1	1
[156]	platform alterations, social norms	0.111	0.5	0.333	0.5	0.333	0.33
[185]	misinformation detection, warning labels	0.067	0.1	0.25	0.25	0.333	0
[186]	fake news games, inoculation	0.167	0.25	0.25	0.333	0.25	0.333
[208]	source credibility labels, warning labels	0.077	0.125	0	0	0	0.333
[262]	friction, platform alterations, warning labels	0.077	0.125	0.2	0.667	0.2	0.667
[280]	account removal	0.125	0.143	1	1	0.5	1
[288]	media literacy	0.091	0.143	0.5	1	1	1
[301]	fact-checking	0.125	0	0.5	1	0.5	1
	Avg. Jaccard Similarity:	0.116	0.218	0.418	0.554	0.443	**0.579**

Table 2.3 summarizes the twenty selected papers, their final labels, and the Jaccard similarity comparing the final labels with each model run. Jaccard similarity is a measure that represents the fraction of items that overlap in two sets, with values closer to 1 indicating greater similarity or agreement [139]. The ordered combined prompt demonstrated the highest average performance, with an overall average Jaccard similarity of 0.443 for the 4o-mini model and 0.579 for the 4o model, respectively. Given the substantial improvement provided by the 4o model over the 4o-mini model, the 4o model was employed to label interventions for the rest of the papers.

2.2.6 Label Agreement

Approximately 25% of the included Stage 1 papers were randomly assigned to have two human labelers (101 papers). The remaining papers (265) were assigned to one human labeler and ChatGPT as the second labeler. There were five unique raters: three interns, myself, and ChatGPT. A collaborator, Peter Carragher, and I resolved disagreements between raters to determine the final assigned labels.

Given the varying number of labelers and potential labels per paper, the 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.

Agreement with the final labels varied by rater but was consistently high. Table 2.4 displays the Jaccard similarity agreement with the final label categorized by labeler type (ChatGPT, Expert, Intern). On average, ChatGPT matched or surpassed the agreement level of the interns on the intervention labeling task.

Table 2.4: Jaccard Similarity agreement with the final labels broken up by rater type. Rater types are sorted by total number of papers labeled.

Rater	**Papers Labeled**	**Effectiveness**	**Acceptance**	**Interventions**
Interns (3)	390	0.972	0.921	0.667
ChatGPT	264	0.962	0.864	0.746
Expert	76	0.987	0.974	0.938

However, there was variability among the three interns. Table 2.5 shows the Jaccard similarity agreement with the final labels for each individual rater. ChatGPT had the second highest level of agreement with the final intervention labels, though it was slightly behind the interns on both effectiveness and acceptance. Interns are numbered in order of start date and overall internship duration, with Intern 1 starting and being trained first. Intern 3, who stayed on the project the longest, was by far the most accurate of the interns, indicating that experience likely improved their labeling accuracy.

Considering ChatGPT's relatively high level of agreement with the final intervention labels, the ChatGPT prompts were used to assist with labeling the 86 papers added in Stage 2 of the literature review. I reviewed ChatGPT's labels prior to assigning the final labels to these papers.

Table 2.5: Jaccard Similarity agreement with the final labels broken up by specific raters. Raters are sorted by total number of papers labeled.

Rater	Papers Labeled	Effectiveness	Acceptance	Interventions
ChatGPT	264	0.962	0.864	0.746
Intern 3	201	0.970	0.930	0.737
Intern 2	96	0.979	0.948	0.668
Intern 1	93	0.968	0.871	0.515
Expert	76	0.987	0.974	0.938

2.2.7 Data Availability

Here is the list of all reviewed papers and their labels on a public Zotero repository: https://www.zotero.org/groups/5961522/misinformation_interventions.

2.3 Citation Network Analysis

As Figure 2.1 summarizes, 451 papers were labeled following the Stage 1 and Stage 2 literature searches. All papers were imported into Zotero, a citation management tool[3]. The final effectiveness, acceptance, and intervention labels were assigned as tags to each paper. The included papers were exported as a .bib file and loaded into the ORA software [65], where further data cleaning occurred. Authors who may have published under different names or sometimes omitted middle initials on their work were consolidated into a single author entry. Conference proceedings from different years were aggregated into one venue with the conference title for a more straightforward analysis. This process resulted in 1,147 unique authors and 209 unique publication venues from these 412 papers. The ORA software was used to conduct the citation network analysis and produce network visualizations from this set of papers [65]. The additional data attributes, nodesets, and networks are summarized below.

2.3.1 Data Attributes

Each paper's metadata was augmented with citation counts obtained from Semantic Scholar on February 24th, 2025[4]. These citation counts were retrieved using the Zotero Citation Counts Manager plug-in[5]. Discipline information for all 209 venues was sourced from Scimago's journal rank data from 2023[6]. An R script was developed to calculate additional metrics from this data, such as citation counts by author, topic counts, and a summary of the venue disciplines. By far, the most popular venue disciplines were the social sciences (n = 117), computer science (n = 57), and psychology (n = 50). Only 34 venues were exclusively from disciplines that were not among

[3] https://www.zotero.org

[4] https://www.semanticscholar.org

[5] https://github.com/eschnett/zotero-citationcounts [Accessed 02-24-2025]

[6] https://www.scimagojr.com/journalrank.php [Accessed 06-12-2024]

the three main disciplines or a multidisciplinary journal. To reduce the number of discipline tags, any discipline not among the primary three or multidisciplinary was labeled as “other.”

2.3.2 Nodesets

This dataset has four nodesets: Articles, Authors, Publication Venues, and Topics.

- *Article* (n = 451) - The Article nodeset contains the article title, publication year, and DOI extracted directly from the paper’s metadata. Additional attributes pulled into ORA are citation counts (sourced from Semantic Scholar) and whether the paper studies effectiveness (Yes/No) or acceptance (Yes/No).
- *Author* (n = 1147) - The Author nodeset contains the unique author names. It additionally contains the following attributes: citation counts (summed over the Article citation counts for each author), the number of articles per author, and the average number of citations per article for each author.
- *Publication Venue* (n = 209) - The Publication Venue nodeset includes the venue title, the ISSN, and the venue type (journal, conference proceedings, or book). Discipline is an externally sourced attribute from SCImago, representing the primary research areas that the venue publishes. Venues can be classified as belonging to more than one discipline.
- *Topic* (n = 35) - The Topic nodeset includes 35 unique intervention labels. Additional attributes are the topic’s General Intervention Category (see Table 2.1) and Count (the total number of papers associated with each intervention label).

2.3.3 Networks

ORA creates multiple networks from this data. The networks that are analyzed further are described below:

- *Publication Venue x Publication Venue (Co-Authorship)* - This is a symmetric network where a link exists if an author has published in both venues. Link values are weighted and represent the number of authors who have published in both venues.
    - Network Density: 0.0349
    - Link Statistics: 1516 total links, values range from 1-12, mean value is 1.89
    - Component Statistics: 73 isolates, 6 dyads, 2 triads, 2 larger components
    - Largest Component: 114 authors
- *Topic x Topic (Co-Topic)* - This is a symmetric network where a link exists between two topics if there is an article that covers both topics. Link values are weighted and represent the number of papers that study both intervention topics.
    - Network Density: 0.397
    - Link Statistics: 472 total links, values range from 1-26, mean value is 2.71
    - Component Statistics: 1 isolate (data sharing), 1 large component
    - Largest Component: 34 topics

- *Author x Author (Co-Authorship)* - This is a symmetric network where a link exists between two authors if they have published together. Link values are weighted and indicate the number of papers the authors have co-authored together.
  - Network Density: 0.005
  - Link Statistics: 6654 total links, values range from 1-19, mean value is 1.12
  - Component Statistics: 24 isolates, 53 dyads, 48 triads, 93 larger components
  - Largest Component: 326 authors
- *Publication Venue x Topic* - This network was created by folding the *Publication Venue x Article* and *Article x Topic* networks. The links represent the number of papers published in that venue discussing each intervention topic.
  - Network Density: 0.082
  - Link Statistics: 603 total links, values range from 1-7, mean value is 1.29
  - Component Statistics: 1 large component
  - Largest Component: 244

2.4 Results

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

2.4.1 Descriptive Statistics

The ORA Report "Bibliography and Citation" was run on this dataset. This report provides network metrics and analyzes leading authors, venues, and topics. Figure 2.2 shows the number of papers in this review published each 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 [79].

Next, we investigate the most cited papers overall. Table 2.6 shows the top 10 most cited papers in the dataset. These papers are predominantly in multidisciplinary journals, with only one conference paper making the list.

2.4.2 Publication Venue Analysis

In this section, we examine the leading publication venues in the dataset and visualize the *Publication Venue x Publication Venue (Co-Authorship)* network.

Table 2.7 shows the top 10 publication venues in the *Publication Venue x Publication Venue* network by two centrality 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," or with the most connections to other venues. The second metric is eigenvector centrality, which is another way to measure how influential a node is in a network. A venue has high eigenvector centrality if connected to other high-scoring venues.

Table 2.6: The top 10 most cited papers in the data set.

Short Paper Title	**DOI**	**Year**	**Venue**	**Discipline**	**Citations**
1. Information Credibility on Twitter	10.1145/ 1963405.1963500	2011	*Web Conference*	Computer Science	2370
2. When Corrections Fail: The Persistence of Political Misperceptions	10.1007/ s11109-010-9112-2	2010	*Political Behavior*	Social Sciences	2317
3. Fighting COVID-19 Misinformation on Social Media	10.1177/ 0956797620939054	2020	*Psychological Science*	Psychology	1280
4. Effective Messages in Vaccine Promotion: A Randomized Trial	10.1542/ peds.2013-2365	2014	*Pediatrics*	Medicine	1089
5. Prior exposure increases perceived accuracy of fake news	10.1037/ xge0000465	2018	*Journal of Experimental Psychology: General*	Psychology	831
6. Inoculating the Public against Misinformation about Climate Change	10.1002/ gch2.201600008	2017	*Global Challenges*	Multidisciplinary	674
7. Shifting attention to accuracy can reduce misinformation online	10.1038/ s41586-021-03344-2	2021	*Nature*	Multidisciplinary	622
8. Analysing How People Orient to and Spread Rumours in Social Media	10.1371/ journal.pone.0150989	2016	*PlOS ONE*	Multidisciplinary	622
9. Fighting misinformation on social media using crowdsourced judgments of news source quality	10.1073/ pnas.1806781116	2019	*PNAS*	Multidisciplinary	594
10. Neutralizing misinformation through inoculation	10.1371/ journal.pone.0175799	2017	*PLOS ONE*	Multidisciplinary	592

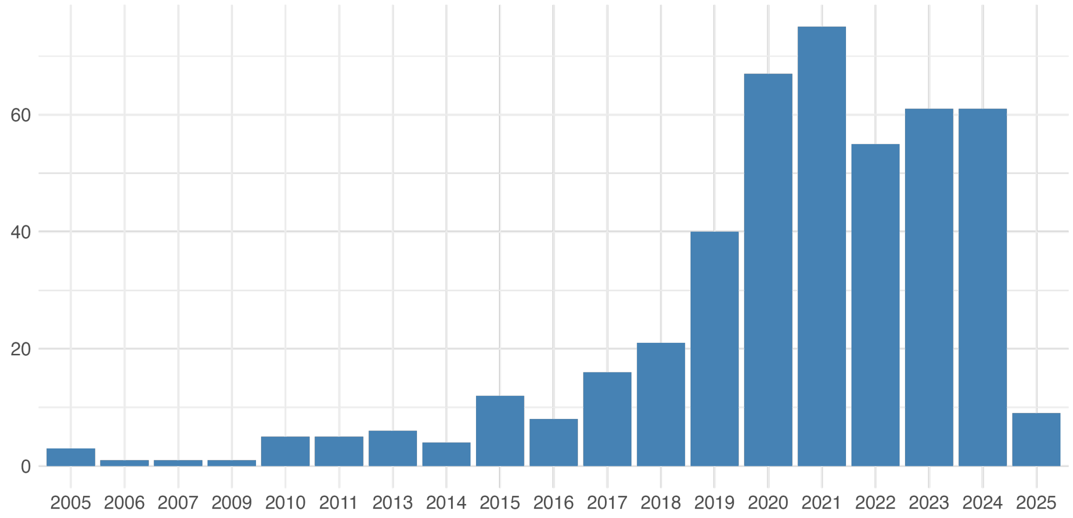


Figure 2.2: Number of articles in this review published per year.

Table 2.7: The top 10 venues in the Co-Publication Venue network.

Rank	Total Degree Centrality	Eigenvector Centrality
1.	Nature Human Behaviour	Nature Human Behaviour
2.	Harvard Misinformation Review	Science Advances
3.	Science Advances	Harvard Misinformation Review
4.	Royal Society Open Science	Psychological Science
5.	PNAS	PNAS
6.	Psychological Science	Royal Society Open Science
7.	J. of Experimental Psych: General	J. of Experimental Psych: General
8.	Journal of Communication	Nature
9.	CSCW; J. of Applied Social Psych.; Political Psych.	CSCW
10.	Cognitive Research: Principles and Implications	J. of Applied Social Pscyh.

Table 2.8 ranks the venues according to the number of publications in this data set. This table demonstrates that having many publications does not necessarily ensure a top ranking for that venue based on centrality metrics. Several of these venues, including *Scientific Reports* and the *CHI Conference*, do not rank highly in terms of network centrality (Table 2.7).

Table 2.8: The top venues by total number of publications in the data set.

Publication Venue	**Count**	**Discipline**
1. CSCW Conference	18	Computer Science
2. Harvard Misinfo Review	15	Social Sciences
3. PLOS ONE	11	Multidisciplinary
4. CHI Conference	9	Computer Science
4. Science Communication	9	Social Sciences
5. Scientific Reports	8	Multidisciplinary
6. Journal of Communication	7	Social Sciences
6. Nature Human Behaviour	7	Psychology, Other
7. Cognitive Research: Principles and Implications	6	Psychology, Other
7. Health Communication	6	Social Sciences
7. J. of Applied Research in Memory & Cognition	6	Psychology
7. Memory & Cognition	6	Psychology, Other
7. Political Behavior	6	Social Sciences
7. PNAS	6	Multidisciplinary

Next, we analyze the Co-Publication Venue network. Figure 2.3 illustrates the largest component. Nodes are sized by total degree centrality and colored by discipline. A selection of venues is highlighted. The density of this network is low, at 0.0349. 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 both total degree centrality and eigenvector 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 2.4 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 place. 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.

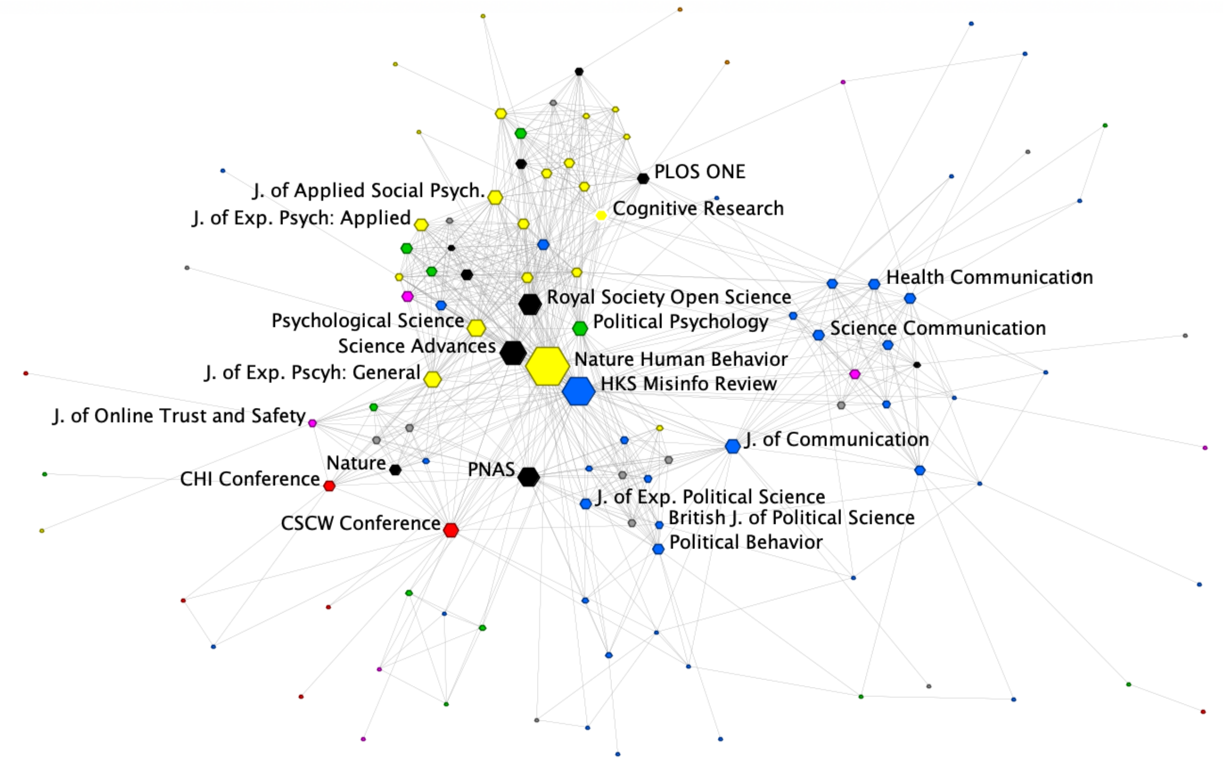


Figure 2.3: Co-Publication Venue network. Nodes are sized by total degree centrality and colored by discipline (red for computer science, yellow for psychology, blue for social sciences, black for multidisciplinary, and grey for other.)

.

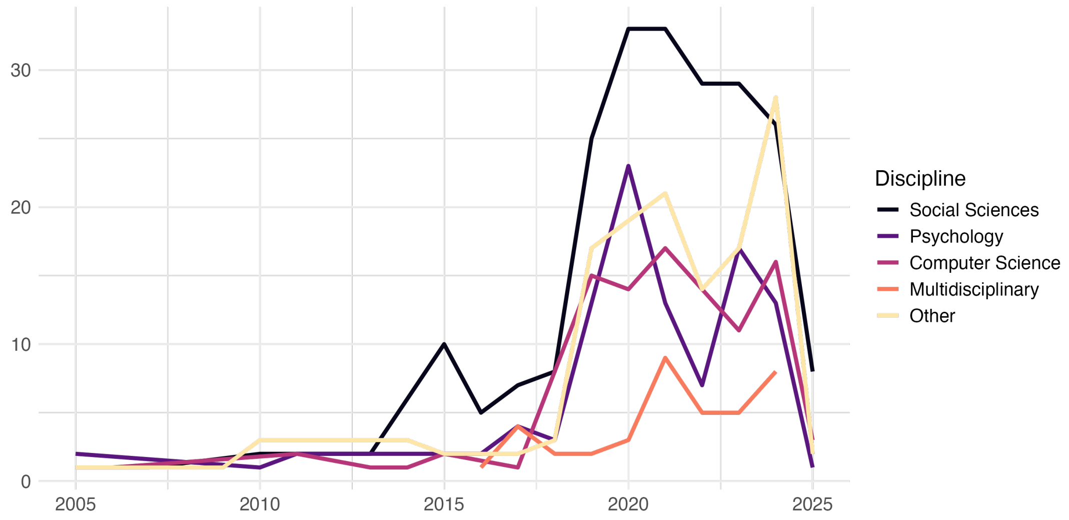


Figure 2.4: Number of articles in this review by venue discipline.

2.4.3 Topic Analysis

In this section, we analyze the leading topics, investigating whether certain interventions are over-studied or under-studied, and identifying topics that are often studied together.

First, we calculate the descriptive statistics for the number of papers assigned to each label. Table 2.9 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 studied interventions are featured in fewer than 10 papers.

The summary statistics for both paper and author counts are presented in Table 2.10 and further underscore this discrepancy. 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, we found that 404 papers (89.6%) analyzed an intervention's effectiveness, and 40 papers analyzed an intervention's level of user acceptance (8.9%) (See Table 2.11).

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 2.5 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 2.5, 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. The overstudied and understudied topics align with Courchesne et al.'s previous review article. However, their metric for categorizing over and understudied topics was based on what the platforms were actually implementing [79].

Next, we analyze whether the types of interventions studied have changed over time. Figure 2.6 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 interventions 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.

2.4.4 Author Analysis

In this section, we analyze the top authors in the data and visualize the Co-Authorship network. Table 2.12 displays the top 10 most cited authors, including only those who have published more

Table 2.9: 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 content 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

Table 2.10: Statistical summary of the number of papers and unique authors by intervention type. Outliers are identified from the inter-quartile range.

	Mean (SD)	1st Q.	Median	3rd Q.	Outliers
Papers per Intervention	23.2 (30.7)	5	12	28	fact-checking, debunking, media literacy
Authors per Intervention	76.3 (92.4)	13	41	81.5	fact-checking, debunking, media literacy, inoculation

Table 2.11: The number of studies examining the effectiveness or acceptance of one or more interventions.

		Studies Acceptance	
		Yes	No
Studies Effectiveness	Yes	24 (5.3%)	380 (84.3%)
	No	16 (3.5%)	31 (6.9%)

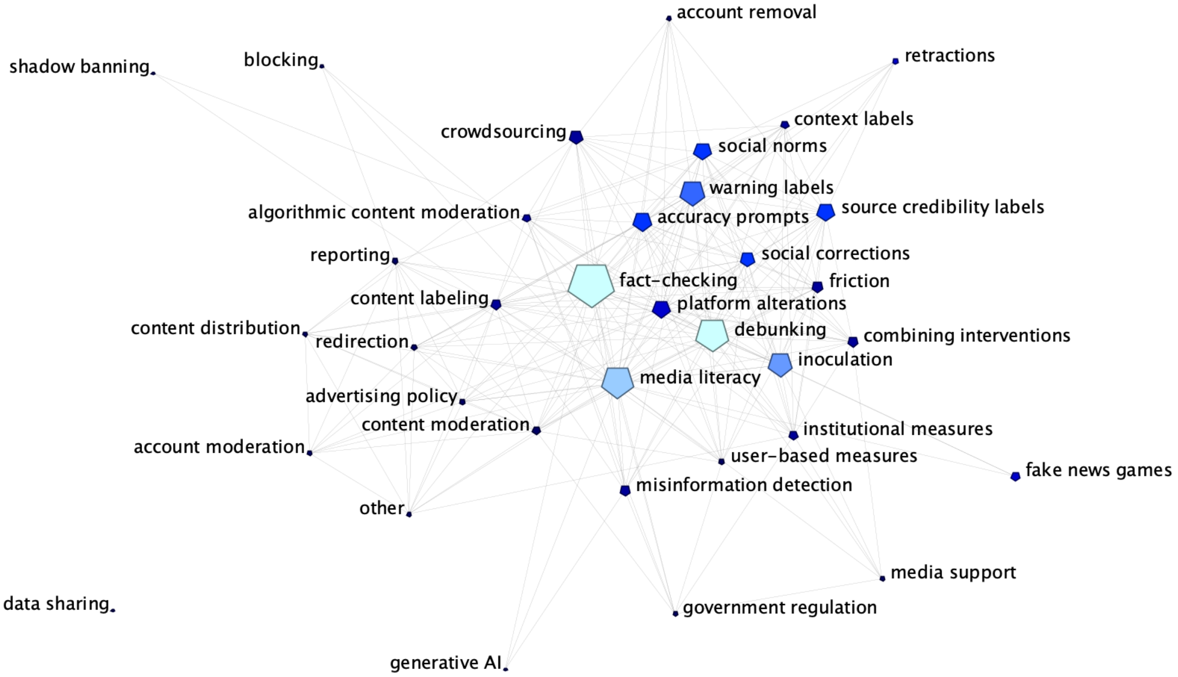


Figure 2.5: Co-Topic network. Nodes are sized by total degree centrality and colored by paper count. The lighter the blue, the more papers that study that topic.

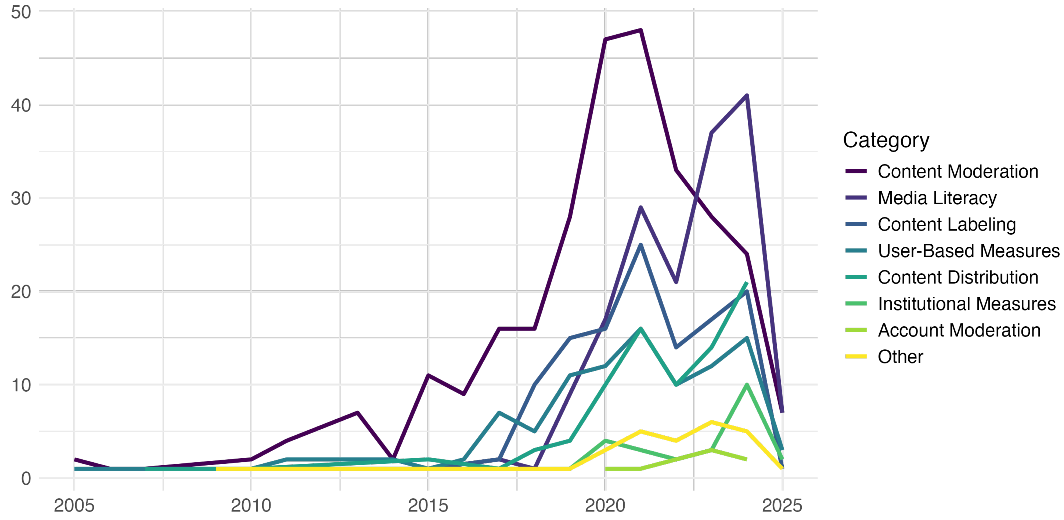


Figure 2.6: Number of articles in this review by general intervention category studied.

than once in this dataset. The vast majority of authors (83% or 953 authors) have only one paper in this dataset. These ten authors are also the same authors who have the most papers in the dataset.

Table 2.12: The top 10 most cited authors with more than one paper in the data set.

Author Name	Citation Count	Article Count
1. Brendan Nyhan	5549	17
2. Jason Reifler	5305	12
3. David G. Rand	5102	22
4. Gordon Pennycook	5004	21
5. Sander van der Linden	3122	22
6. Ullrich K. H. Ecker	2981	30
7. Stephan Lewandowsky	2735	22
8. Emily K. Vraga	2332	16
9. Leticia Bode	2178	13
10. Jon Roozenbeek	2171	16

Table 2.13 shows the top 10 authors in the *Author x Author (Co-Authorship)* network based on three metrics: total degree centrality, eigenvector centrality, and betweenness centrality. Betweenness centrality is a metric that identifies bridging nodes by quantifying the number of shortest paths that pass through each node. Authors with high betweenness centrality are potentially influential because they can facilitate connections between authors who are otherwise unconnected or have not previously collaborated.

Next, we visualize the *Author x Author (Co-Authorship)* network in Figure 2.7. 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

Table 2.13: The top 10 authors in the Co-Authorship network.

Rank	Total Degree Centrality	Eigenvector Centrality	Betweenness Centrality
1.	Stephan Lewandowksy	Stephan Lewandowksy	Brendan Nyhan
2.	Sander van der Linden	Gordon Pennycook	Jason Reifler
3.	Ullrich K. H. Ecker	David G. Rand	Ullrich K. H. Ecker
4.	Gordon Pennycook	Ullrich K. H. Ecker	Andrew M. Guess
5.	David G. Rand	Sander van der Linden	Sander van der Linden
6.	Brendan Nyhan	Adam J. Berinsky	Gordon Pennycook
7.	Adam J. Berinsky	Briony Swire-Thompson	David G. Rand
8.	Philipp Schmid	Rakoen Maertens	Stephan Lewandowksy
9.	Cornelia Betsch	Philipp Schmid	John Cook
10.	Jason Reifler	Melisa Basol	Adam J. Berinsky

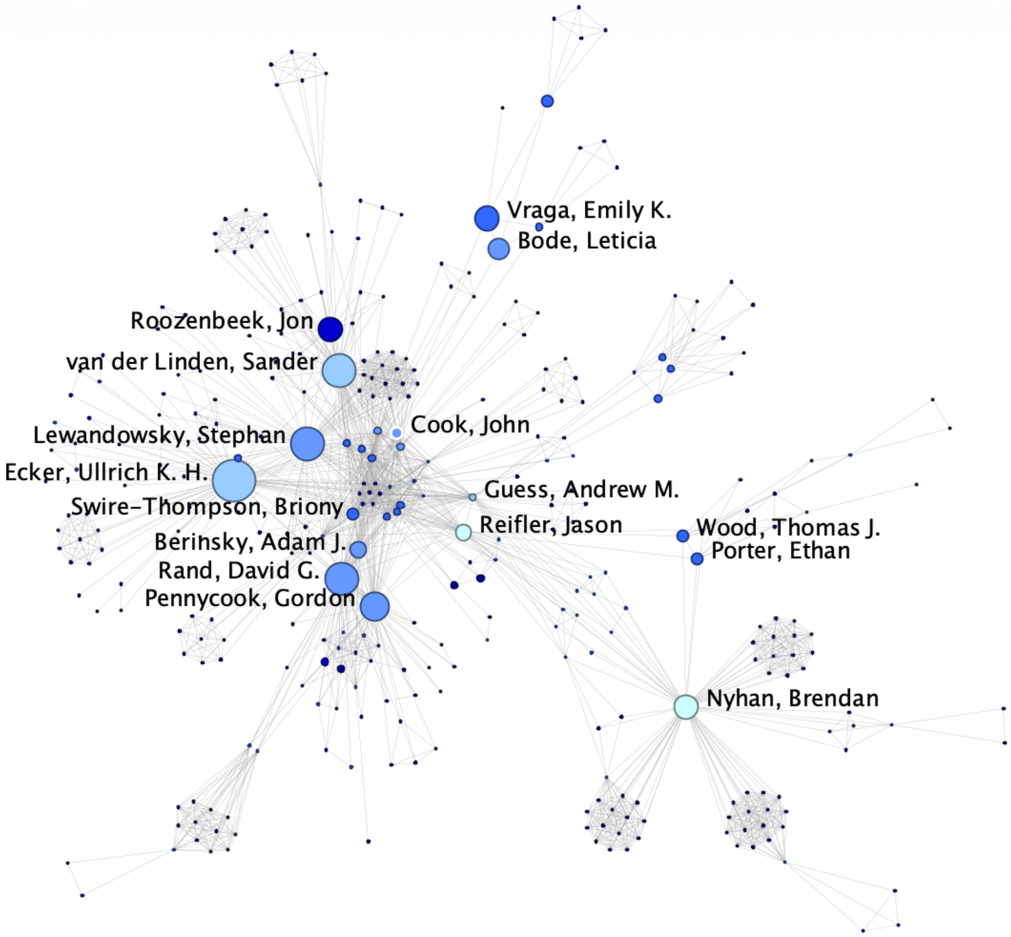


Figure 2.7: The largest component of the Co-Authorship network. Nodes are sized by total number of papers and colored by betweenness centrality, with lighter blue colors indicating higher betweenness.

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 countermeasures 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 [167], 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 (Table 2.12). 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 2.8. 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.

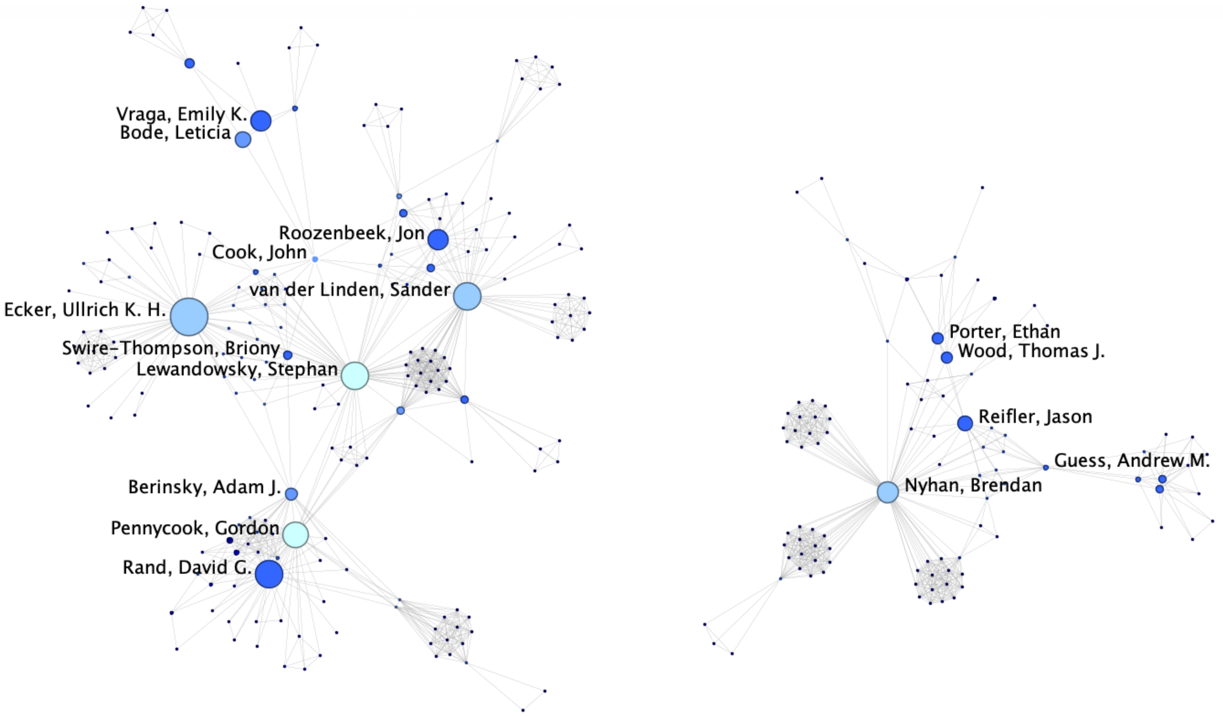


Figure 2.8: The largest component of the Co-Authorship network excluding one paper with 30 authors. Nodes are sized by total number of papers and colored by betweenness centrality, with lighter blue colors indicating higher betweenness.

2.5 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, indicating that this is a growing field of literature.

First, we analyzed the publication venues in this dataset. The Co-Publication Venue network reveals disjointedness in the literature, with most venues clustered near others within the same discipline. While there are several journals, such as the *Harvard Misinformation Review* and *Nature Human Behavior*, that 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, Journalism and Communication, and Computer Science. This research area is highly interdisciplinary.

Next, we analyzed the top topics studied in the literature. User acceptance is largely overlooked in the literature, 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 [184]. It is also notable that many studies on intervention effectiveness employ survey instruments which could be easily extended to measure user acceptance [27].

Furthermore, as found in a previous review article [79], 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 [315]. 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 [131], 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 user account demonetization strategies. While platforms do actively demonetize accounts that spread misinformation[7] [8], removing their ability to make money from posted content, a lack of empirical research on this issue is worrying given the financial incentives behind misinformation such as running ads and selling merchandise [190, 221]. Only one study addressed interventions that target the viability of, or ‘cost’ of, propagating misinformation on social media [138]. As monetization interventions require platform access to advertising and payment data, greater collaboration between academic institutions and social media platforms is required to tackle monetization.

Finally, we encountered disagreements on the effectiveness of certain interventions while analyzing the literature. For example, amongst the most studied countermeasures, we find several sources of contention; a body of work claims the effectiveness of the “Bad News” game for inoculation [36, 250], while a meta-review finds their results to be insignificant using pre & post treatment classification accuracy [36]. The efficacy of contextual labels, such as Community Notes, is also unclear, with recent studies indicating mixed effects [46], or no effect [71], on misinformation exposures. The effectiveness of source credibility interventions has also been

[7] https://www.youtube.com/howyoutubeworks/our-commitments/fighting-misinformation/#rewarding-trusted-creators-and-artists [Accessed 04-14-2025]

[8] https://www.tiktok.com/transparency/en-us/combating-misinformation/ [Accessed 04-14-2025]

disputed [58]. 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 [27, 121].

2.6 Conclusions

2.6.1 Limitations

One of the main concerns for any study of countermeasures to misinformation is that the method for selecting papers may have missed articles in specific sub-areas. In this analysis, 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. Lastly, only articles written in English were included in our analysis. However, the categories of misinformation identified as over-studied and under-studied are similar to those that other related review articles found, mitigating this risk [79].

Another limitation is that we did not have a baseline against which to compare this bibliometric analysis. For example, is it typical for co-authorship networks to fragment as much as they did in this work when one article with a high number of authors is removed (see Figure 2.7 and 2.8)? Is the clustering of journals by discipline common in other multidisciplinary fields? These questions suggest possible future directions for this work.

2.6.2 Contributions

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, including user-based countermeasures, which will be examined further in Chapters 3 and 4. Additionally, we found that the academic literature primarily focuses on the effectiveness of countermeasures without addressing the critical metric of user acceptance. If user acceptance is low, implementing that intervention is unlikely, making its relative effectiveness less relevant. User acceptance is the main aspect studied in the subsequent chapters of this dissertation.

The analysis also uncovered structural issues in the research ecosystem. Publication venues in this field have primarily remained clustered by discipline, although several collaborative and multidisciplinary publications have emerged in recent years. Despite this, the field remains fragmented: there is limited integration across intervention types, inconsistent findings on effectiveness, and a lack of cohesion in the authorship network. Future work should evaluate interventions not only in terms of effectiveness but also in terms of acceptance and other relevant factors. The lack of consensus on various interventions emphasizes the need for comprehensive evaluation metrics [291] in the field and highlights the importance of meta-reviews. Chapter 6 will review the effectiveness and acceptance of interventions in greater detail.

Chapter 3

Characterizing User-based Countermeasures

In Chapter 2, I demonstrated that user-based interventions are an understudied type of countermeasure for combatting misinformation. Examining individual behavior in response to encountering misinformation is crucial because previous research has shown that debunking myths is more effective when it comes from a trusted source, such as a friend or family member [182]. Additionally, social media corrections have been shown to be highly effective [28, 48, 303]. This suggests that individuals responding directly to misinformation in real-time can help slow or stop its spread.

In this chapter, I investigate the behavior and opinions of social media users when they see or post misinformation. We surveyed 1,010 active social media users residing in the United States. This survey covered the social media platforms where they encounter misinformation, if they have posted misinformation unintentionally, their reactions to seeing or posting misinformation, and their opinions on how they think others should act.

The primary research question for this chapter is: How are people currently tackling misinformation on social media? More specifically,

1. How do people respond to misinformation posted by others or themselves, and how do they think others should respond?
2. Do people respond differently depending on who posted the misinformation **(poster)** and where it was posted **(platform)**?
3. What demographic factors, if any, are associated with opinions on these topics?

3.1 Introduction

As described in Chapter 2, many countermeasures, including user-based countermeasures, have been understudied in the literature to date [79]. Recent research has begun investigating the relationship between seeing misinformation countermeasures online and public perception of those countermeasures [255]. However, the experience of observing or posting misinformation differs from the experience of observing or conducting a countermeasure. Viewers of misinformation on social media can either directly confront the authors of misinformation by offering social corrections or indirectly counter the misinformation by, for example, reporting the misinformation. 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 [48, 182]. This suggests that individuals responding directly to misinformation from users in their network can help slow or even stop the spread of misinformation.

This work investigates if these social corrections happen and if they depend on the nature of the relationship between the misinformation poster and the observer or the platform on which it was posted. Given the scale in which social media companies must detect and respond to misinformation [152], users can play a crucial role in limiting the spread in real time. Indeed, social media companies such as X (formerly known as Twitter) have piloted programs like Community Notes where users can add corrections and/or context to tweets they deem misleading [299]. Additionally, previous research has shown that most people do not intend to spread misinformation [19, 35, 307]. Instead of consciously sharing misinformation, cognitive and socio-affective mechanisms (e.g., intuitive thinking, identity motives) facilitate sharing and even belief in misinformation in some cases [69, 97]. If this is the case, nudging social media users to focus on accuracy goals could help limit the unintentional spread of misinformation [231].

In addition, we explore what people believe others should do when they see misinformation or post misinformation themselves. This provides researchers and policymakers a sense of what social media users want the norm response to be, which is essential for public outreach about crowdsourced misinformation mitigation. We also examine how expectations vary from reported actions to understand the extent to which people currently feel empowered to respond to misinformation regardless of (their own) situational constraints. Participants may want others to respond to misinformation with higher effort actions than they do themselves. This act of hypocrisy, failing to practice what one preaches [59], could be leveraged to induce prosocial behavior changes (e.g., directly addressing content they believe contains misinformation) [20, 103, 270]. Making the discrepancy between behavior and advocated norms for behavior salient can activate threats to self-integrity, driving behavior in line with advocated norms to minimize the dissonance [270].

This study surveyed 1,010 United States residents who use social media at least weekly. This survey covered the social media platforms where participants encounter misinformation, if they have posted misinformation (intentionally or unintentionally), their response to seeing or posting misinformation, and their opinions on how they think others should respond to seeing or posting misinformation. Participants report their response and the response they expect others to do when seeing misinformation posted by someone else for three levels of closeness to the sender of misinformation: close (e.g., a close friend or family member), somewhat close (e.g., acquaintances, colleagues, friends, extended family), or not close (e.g., someone you do not know offline). We included seven ways to counter misinformation posted by others, four ways to respond to misinformation posted by oneself, and an option to engage in no action. These responses capture a broad range of actions that involve directly and indirectly interacting with the content containing misinformation.

This chapter is divided into two main analysis sections: **closeness analysis** and **platform analysis**. The research questions, hypotheses, and analysis plan associated with the closeness analysis were pre-registered and published at *Scientific Reports*[159].

3.2 Data and Methods

3.2.1 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 did not collect personally identifiable information. Informed consent was obtained from all participants. We expected the survey to take 15-18 minutes based on pilot tests. Participants were paid $3 each, which is equivalent to $10/hour if they took 18 minutes to complete the survey.

3.2.2 Pilot Data

The survey was implemented in Qualtrics and was sent out to a small sample of Cloud Research Mechanical Turk participants to ensure questions were straightforward and the bot and duplicate detection worked. Twenty-two participants attempted the survey: 14 were excluded, most of them automatically by Cloud Research, for either being spam/bots, being a duplicate response, or failing to pass the screening questions (18+, U.S. resident, use social media weekly). Participants excluded for these reasons were removed at the beginning of the survey and were not paid.

Participants were also asked if they had any comments. As a result, a few questions were removed to prevent the survey from being too long or reworded for succinctness and clarity. This document's hypotheses and research questions are based on the revised survey. While a pilot sample of eight responses is small, it helped improve the research design. It demonstrated that the questions were understandable and that this survey could effectively address the hypotheses and research questions. See the Supplementary Information in the Stage 1 Protocol[1] for more detailed information on the pilot data.

3.2.3 Survey Design and Sampling Plan

This section describes our survey design, sample characteristics, sample size determination, data exclusion criteria, and all primary measures. The survey was designed to answer the research questions associated with both Chapters 3 and 5. See `https://doi.org/10.1184/R1/27264813` to see a copy of the survey.

Participants

Our survey had 1,010 participants, and the data was collected between July and August 2024. This sample size was deemed appropriate because it provided sufficient power for our proposed hypotheses. The survey was implemented using Qualtrics and administered through Cloud Research, an online recruiting platform [126], using Mechanical Turk survey participants. Only those respondents who are United States residents, adults, and use social media at least once a week were given the entire survey.

[1] `https://figshare.com/s/683b1e7c2f2bad96f604`

Procedure

Qualifying Questions: We employed several methods to recruit relevant participants and maintain high data quality. Participants were adult U.S. residents who use social media weekly, and they also must have met the following criteria for inclusion:

1. Approved by Cloud Research
2. Had a higher than 95% approval rating on Mechanical Turk
3. Finished the survey
4. Not a bot (both Qualtrics and Cloud Research have bot detection) [1]
5. Not a duplicate response (Qualtrics flags likely duplicate responses) [1]

Data that failed even one of these criteria was excluded. A question at the end of the study asked if participants answered randomly at any point. This measure was for data quality purposes only and was not used to exclude data. 997 participants (99.7%) responded with "no", while 3 participants responded with "yes" (0.3%), and 10 skipped the question. The median time to complete the survey was 11.0 minutes, while the mean was 14.3 minutes.

Previous research has suggested that Mechanical Turks' data quality is high and that Turkers are more likely to pass attention-checking questions than other online panels [125]. A previous study also shows that the 95% approval rate cut-off can ensure high-quality data without using attention-checking questions [224]. Therefore, this survey did not have any attention-checking questions.

Behavioral Questions: This section asked participants how they respond to seeing misinformation on social media platforms they use and how they react if they realize they have posted misinformation.

First, participants were asked if any of their social media contacts have ever posted something they believe to be misinformation, how often they saw it, and on which platforms they saw it. They were able to select among the top 11 most frequently used platforms in the United States as determined by Pew Research [25]. They also had the option to write in another platform that was not listed. Then, for all platforms they claimed to have seen misinformation on, they were asked how close they were to the people posting misinformation and how they responded. Possible responses are shown in Table 3.1.

Next, participants were asked if they had ever intentionally or unintentionally posted misinformation. If they have unintentionally posted misinformation, they were asked on which platforms and then asked what they did on each platform once they realized they posted misinformation. Possible actions they could have taken are shown in Table 3.2.

Belief Questions: This section asked participants how they thought people should respond when seeing misinformation. The questions were broken up by closeness: how should people respond to misinformation posted by a close contact? A somewhat close contact? A not close contact? Again, possible responses they were able to select are described in Table 3.1. Finally, participants were asked what people should do if they realize they have posted misinformation (possible actions are described in Table 3.2).

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

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

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

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

Demographic Questions: This section asked participants for various demographic characteristics. These were age, gender, race, ethnicity, education, household income, religion, political party affiliation, and general political leanings.

3.2.4 Measures

We created measures to quantify the amount of effort put into responding to misinformation on social media. For both the closeness and platform analysis measures, anything labeled in Tables 3.1 or 3.2 as no effort received a score of 0, low effort received a score of 1, and high effort a score of 2. Anything without a label was labeled as NA. Participants were given the value of the highest effort level they engaged in per closeness level or per platform.

Closeness Analysis Measures

Measure 1a): Effort Expended to Respond to Misinformation Posted by Others based on Closeness (Behavior)

Measure 1b): Effort Expended to Respond to Misinformation Posted by Others based on Closeness (Opinion)

There were three calculated values for both measures, one per closeness level. Table 3.1 shows a list of possible responses one could have when seeing misinformation on social media, generalized to apply to various social media platforms, and rated as no effort, low effort, or high effort. The only no-effort response is ignoring the post. Respondents could also respond with "I don't remember," in which case their effort level was not recorded, as it is unknown. A low-effort response means an action was taken, but there was no interaction with the content directly. A high-effort response indicates that the user likely took more time to respond and interacted with the content directly. The participants were able to select more than one of these actions. Pilot data values for Measures 1a) and 1b) for somewhat close contacts are in the Supplementary Information file in the Stage 1 Protocol[1].

Measure 2a): Effort Expended to Respond to Misinformation Posted by Oneself (Behavior)

Measure 2b): Effort Expended to Respond to Misinformation Posted by Oneself (Opinion)

We created measures 2a) and 2b) to quantify the effort one puts into correcting misinformation they posted online. Table 3.2 shows a list of possible actions someone could take. They are rated in the same way as the efforts described in Table 3.1: no, low, or high effort. Like in Table 3.1, the only no-effort response is leaving the post as is. Respondents could also respond with "I don't remember," in which case their effort level was not recorded, as it is unknown. Deleting the post is categorized as low effort. The remaining actions are classified as high effort, as they indicate the user took more time to respond and they placed effort into correcting their mistake. Pilot data values for Measures 2a) and 2b) are in the Supplementary Information file in the Stage 1 Protocol[1].

Platform Analysis Measures

Measure 3: Effort Expended to Respond to Misinformation Posted by Others based on Platform (Behavior)

Measure 4: Effort Expended to Respond to Misinformation Posted by Oneself based on Platform (Behavior)

Similarly, for the platform analysis we calculate the effort one puts into correcting others or themselves, but broken up by platform instead of closeness. Opinion questions were not asked based on platform, so measures for those are not included here.

3.3 Closeness Analysis

We analyzed closeness to misinformation poster in our Registered Report at *Scientific Reports*. This study involved the following research questions. See the Design Table in Appendix C or the Stage 1 protocol[1] for more details.

1. How do people respond and think others should respond when they see misinformation? Do response(s) change based on how close the participant is to the poster of misinformation?
2. How do people respond and think others should respond when they realize they have posted misinformation?

3. How do people's responses and beliefs about how others should respond after seeing misinformation differ from their responses and beliefs when they realize they have posted misinformation?
4. How do beliefs about responses to misinformation differ based on various demographic factors?

3.3.1 Related Work

Social media users may refrain from directly responding to (recognized) misinformation due to a myriad of constraints, such as concerns over damaging interpersonal relationships or their own credibility [274]. In addition, verifying and correcting misinformative claims is a time-consuming, effortful process in practice [267]. Users may also feel helpless to counter misinformation given the vast amount available online [274]. We generally expect people will incorporate these constraints more when reporting their own response to misinformation than when describing expectations for others due to cognitive distortions like fundamental attribution error [171, 278]. While users can account for the factors that drive their decision-making about how to respond to misinformation online, it is significantly more challenging to incorporate the hypothetical situational constraints of others. Moreover, asserting that others should respond with high levels of effort can uphold feelings of morality even if participants do not want to engage in the moral behavior (i.e., responding actively to misinformation online) for whatever reason [38, 170]. Therefore, we have the following two hypotheses:

H1.1: *People believe individuals should expend more effort to respond to misinformation online than they actually do.*

H2.1: *People believe others should expend more effort to respond to misinformation online after realizing they posted misinformation than what they actually do.*

Previous work shows users are more likely to correct a close contact because it is perceived as more worthwhile [274]. If users are going to take the time to engage with misinformative content directly, they want to feel like it will have an impact. If they have a personal relationship with the sender of misinformation, they have more information about the expected effectiveness of their correction. Furthermore, people may be especially concerned about close contacts believing in misinformation due to the potential negative consequences. We expect this will translate to expectations for others as well:

H1.2: *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.*

H1.3: *People believe others should 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.*

When comparing responses to misinformation posted by others and posted by oneself, many individual and social factors may come into play, including wanting to preserve harmony and credibility or avoid embarrassment. Previous research from Singapore has shown that many young people avoid correcting misinformation posted by others to maintain their interpersonal relationships but do correct themselves to preserve their credibility despite possible embarrassment [213]. Other work suggests that many only correct others if they are close contacts or when it is an issue they care about [274]. Due to the conflicting literature in this area, we have created two non-directional hypotheses where we expect that people respond and expect others to respond with a different level of effort when the sender of misinformation is someone else compared to themselves:

H3.1: *People respond with a different level of effort when the sender of misinformation is someone else compared to themselves.*

H3.2: *People want others to respond with a different level of effort when the sender of misinformation is someone else compared to themselves.*

Finally, we investigate how behavior and beliefs about responses to misinformation on social media vary by partisanship and other demographic factors (**RQ4**). Extensive previous research has examined differences in misinformation susceptibility across age, gender, education level, income bracket, religious groups [57, 64, 91, 207, 293], and partisan groups [107, 114], as well as the effectiveness of interventions across demographic groups [122]. Increasingly, researchers are studying how individual factors impact support for misinformation interventions, typically at a platform level [93, 160, 165, 255]. In specific contexts, such as highly partisan environments, less susceptibility to misinformation is associated with more support for platform interventions. Left-leaning, Democratic individuals are both more supportive of platform interventions [166, 199, 255], and less likely to observe or spread misinformation online [107, 114], than right-leaning or Republican individuals. In other cases, high susceptibility is linked to more support. Older adults are typically associated with higher susceptibility to sharing and believing misinformation [55, 120]. Yet there is evidence that they support nudge interventions more [160].

Crucially, the type of misinformation seems to substantially affect susceptibility and views of countermeasures (e.g., older adults seem less susceptible to health-related misinformation than younger people [207]). Given the variability in susceptibility and support for platform-level interventions, we conducted exploratory analyses on how demographic attributes and partisanship affect support for and engagement with individual interventions. Understanding how specific populations, particularly those shown to be highly susceptible to misinformation, view and enact individual interventions informs effective public messaging about countering harmful content online.

3.3.2 Bayesian Power Analysis

We used a Bayesian approach to test our hypotheses. Unlike frequentist methods and p-values, the Bayes factor can show evidence in favor of either the null or the alternate, not just “reject” or “fail to reject” [169, 260]. Additionally, unlike a traditional confidence interval that gives the range of values that would not be rejected at a specified p-value, the highest density interval includes, say, the 95% high probable values for the estimated parameter [169]. Finally, corrections are typically needed for multiple t-tests due to a concern over the possible detection of false positives, and these corrections can result in reduced power [108]. However, Bayesian tests typically do not need a correction for multiple tests, as the Bayesian prior places a relatively high probability on null effects [210].

This work used a Bayes Factor fixed-n design, where n is our sample size. Given our budget, our maximum sample size was determined to be approximately 1000 respondents. We used the methods described in the Schönbrodt and Wagenmakers (2018) design analysis paper [260] and the BFDA R package [259] to run a strength of evidence analysis. Like a traditional power analysis in a classical frequentist approach, a strength of evidence analysis can estimate the sample size needed so that a strong Bayes factor is found in a specific percentage of studies. For this analysis, the Bayes factor threshold was set to 10 and 1/10. Researchers consider a Bayes factor of 10 to be in the "strong" evidence range [175], with 10 meaning that the data is 10 times more likely to follow H1 over H0, and 1/10 indicating the reverse [260]. For this design analysis, evidence with a Bayes factor within that range is deemed "inconclusive." The higher the Bayes factor, the lower the probability of receiving misleading evidence (false positives or false negatives) [260].

We used the JZS Bayes factor, which assumes that the effect of H1 follows a central Cauchy distribution [253, 260]. The JZS Bayes Factor was selected because it is the recommended default when little is known about the expected effect size [253, 260]. The effect size is Cohen's d, the difference in means divided by the standard deviation. The width parameter of the Cauchy distribution was set to $\sqrt{2}/2$, which is the recommended value if expecting smaller effect sizes and is the default used in many software packages, including the R Package BayesFactor [201]. It corresponds to expecting a 50% probability of an effect size between -.707 and .707 for a two-sided test and between 0 and 0.707 for a one-sided test. Pilot data for Hypothesis 1.1 was calculated to have an effect size of 0.80, which we believe is not different enough from the default parameter to warrant changing it, especially with such a small pilot data set.

The strength of evidence analysis was run considering being able to detect a possible effect size of 0.5, which is traditionally interpreted as a "medium" effect size [178]. We chose a medium effect size since we believe a small effect size for our hypotheses would not translate into much practical significance. We used this effect size and a Bayes factor threshold of 10 for the following hypotheses to determine if our sample size was high enough to have a reasonable probability of obtaining strong evidence for an alternate H1. We used the BFDA R package [259], which uses a Monte Carlo method to determine this. It generates 10,000 random samples under H1 with an expected effect size of 0.5 and computes the Bayes factors for those runs given the JZS prior. We ran another 10,000 random samples under H0 and calculated the Bayes factors. Then, we used the distribution of the Bayes factors found to determine the sample size necessary to achieve a high probability (95%) of achieving a Bayes factor of 10.

The following subsections detail the power analysis for each hypothesis. We found that our planned sample size of approximately 1,000 participants was likely sufficient for each hypothesis to detect a medium effect size at a Bayes Factor threshold of 10. The power analysis is additionally summarized in the Design Table in Appendix C.

Hypothesis 1.1

Our first hypothesis is: “Participants say others should respond with higher effort actions to misinformation compared with the actions they take themselves.” Only participants who have seen misinformation on social media will have answers to the part of the survey related to Hypothesis 1.1. It is well-established that people are not always successful or consistent at identifying misinformation online [230]. However, according to an Ipsos survey, of the fraction of American participants that use social media (87%), 77% of them have said they have seen misinformation specifically on social media [264]. In the small pre-test of our survey, 87.5% of participants (7/8) reported having seen misinformation on social media,

We expect almost all of the approximately 1000 participants to have seen misinformation on social media since we are explicitly surveying frequent social media users. Using 77% as a conservative baseline for this power analysis would translate to a sample size of approximately 770 participants. However, not all participants will have seen misinformation at each closeness level, and we do not have an estimate from the literature for how many participants will have seen misinformation from a contact in each closeness category. We estimate that for each closeness level, at least 25% of participants who have seen misinformation will have seen misinformation from a contact at that level, indicating we can achieve a sample size of approximately 192 for this hypothesis.

For Hypothesis 1.1, a one-sided paired Bayesian hypothesis test will be run per closeness level. Running a sample size analysis, we find that to achieve a power of 95%, using a Bayes factor threshold of 10 and effect size of 0.5, the sample size must only be of size 81, which we believe is reachable in each closeness category. The results show that at this sample size under H1, 95.1% of studies correctly identify H1, 4.9% get inconclusive results, and 0% are false negatives. At an estimated sample of 192, we find 100% of runs under H1 find evidence for H1. We additionally find that .2% of runs under H0 find false positive evidence for H1, 40% were inconclusive, and 59.7% correctly showed evidence for H0. To find evidence favoring the null in at least 75% of studies, we would need a sample size of approximately 440 at this high Bayes factor threshold of 10.

Hypothesis 1.2

This one-sided hypothesis states: “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.” As mentioned above under Hypothesis 1.1, we expect most participants to have seen misinformation. Two of the seven individuals who had seen misinformation in our pilot data had seen it at two different closeness levels. Since we do not have an estimate for the fraction of individuals who will have seen misinformation at more than one level, we conservatively estimate that 15% of people out of the at least 77% of people who have seen misinformation will have seen it from more than one closeness type. This translates to approximately 115 participants, greater than the necessary sample size of 81, as described under Hypothesis 1.1.

Hypothesis 1.3

For this hypothesis, all participants will answer the relevant questions, indicating a sample size of approximately 1000. This hypothesis is related to individual opinions on responses to misinformation rather than their actions after seeing it. A one-sided paired Bayesian test will be run for the directional hypothesis. Given such a large sample size of n = 1000, an effect size of 0.5, and a Bayes factor threshold of 10, we find that 100% of the Monte Carlo H1 samples correctly show evidence for H1. Like the other paired directional hypotheses, a sample size of just n = 81 is needed to detect an effect size of 0.5 at a Bayes threshold of 10 with 95% power.

Hypothesis 2.1

This hypothesis is: "People believe others should expend more effort to respond to misinformation online after realizing they posted misinformation than what they actually do." Only those who said they have unintentionally posted misinformation will answer this part of the survey. A Pew Research survey found that 23% of U.S. adults say they have either posted misinformation intentionally or unintentionally [35]. A Statista survey found that 38.2% of Americans claimed they had accidentally shared misinformation on social media [307]. Additionally, almost half of individuals are afraid they may have shared misinformation unintentionally [261]. In our pre-test of 8 participants, we found that half of them said they had posted misinformation accidentally, and a quarter of them admitted to posting misinformation intentionally. Based on the Statista survey, we conservatively estimate that 25% of our participants will say they have posted misinformation accidentally. This is approximately 250 participants. Similar to Hypothesis 1.1, a one-sided paired Bayesian test will be run, and only a sample size of n = 81 is needed to detect an effect size of 0.5 at a Bayes threshold of 10 with 95% power.

Hypothesis 3.1

This two-sided hypothesis is: "People respond with a different level of effort when the sender of misinformation is someone else compared to themselves." Only those who have posted misinformation (we estimate to be at least 77%) and those who have posted misinformation (we estimate 25%) will have data to test this hypothesis. We expect most individuals who have realized they have posted misinformation to have also seen it due to their high awareness of the issue. We conservatively estimate that 15% of all participants will qualify, indicating that the estimated sample size is 150. For a two-sided paired Bayesian test, a sample size of just n = 92 is needed to detect an effect size of 0.5 at a Bayes threshold of 10 with 95% power.

Hypothesis 3.2

Similar to Hypothesis 1.3, all participants will answer the relevant questions because this hypothesis is related to participants' opinions on how to respond after posting misinformation. A two-sided paired Bayesian test will be run, as this hypothesis is non-directional. Given such a large sample size of n = 1000, an effect size of 0.5, and a Bayes factor threshold of 10, we find that 100% of the Monte Carlo H1 samples correctly show evidence for H1. Like Hypothesis 3.1, for a two-sided paired Bayesian test, a sample size of just n = 92 is needed to detect an effect size of 0.5 at a Bayes threshold of 10 with 95% power.

3.3.3 Statistical Methods

In this section, we detail all pre-registered analyses per hypothesis. We ran paired hypothesis tests, treating the ordinal data for effort level as interval data. The literature is divided on whether treating ordinal data as interval is recommended [134, 247, 312]. Given the large sample size, we expect the interpretation will likely not change if we analyze the data as ordinal rather than interval. However, as supplemental work, we conducted robustness checks by analyzing a Chi-square test of independence (treating the data as categorical) to improve the robustness of our results. We performed all statistical tests using the BayesFactor R package [201] and the most recent R and R Studio versions at the time of the analysis (R v4.4.172, RStudio v 2024.04.2+76473).

Hypothesis 1.1

To test **H1.1**, we ran a one-sided paired Bayesian null hypothesis test comparing the effort level of participants' actions when seeing misinformation (Measure 1a) with the effort level participants say others should do when seeing misinformation (Measure 1b) for each closeness level. To run this analysis, we took each user's maximum effort level, ranging from 0-2, for each closeness level. The effect size is equal to (the mean of Measure 1b - mean of Measure 1a) / standard deviation. The null hypothesis (H0) is that the effect size is $<=$ 0 (the null interval range is -Inf to 0). The alternate hypothesis (H1) is that there is a difference in the means with an effect size of greater than 0. The 95% highest density intervals were also calculated.

Hypothesis 2.1

We tested this hypothesis in a similar manner as Hypothesis 1.1. We ran a one-sided paired Bayesian null hypothesis test comparing the effort level of participants' actions when realizing they have posted misinformation (Measure 2a) with the effort level participants say others should do when they realize they have posted misinformation (Measure 2b) for each closeness level. To run this analysis, we took each user's maximum effort level, ranging from 0-2, on each closeness level. We again calculated the 95% highest density interval.

Hypothesis 1.2 and 1.3

Similarly, we ran one-sided Bayesian hypothesis tests for these hypotheses, again using a null interval range of -Inf to 0. The highest density intervals were calculated.

Hypothesis 3.1 and 3.2

For Hypothesis 3.1, we ran a two-sided paired Bayesian hypothesis test comparing the effort level of participants' actions when seeing others post misinformation (Measure 1a) with the effort level of participants' actions when they realized they posted misinformation themselves (Measure 2a). Similarly, for Hypothesis 3.2, we compared Measure 1b and Measure 2b. For both, the null interval range was -0.2 to 0.2. We added a buffer of 0.2 because we consider an effect size that small to be practically equivalent to no effect size. We additionally calculated the highest density interval and made appropriate visualizations for all the results.

3.3.4 Results

We surveyed 1,010 active social media users in the United States. Almost all the participants said they had seen misinformation on at least one social media platform. Table 3.3 shows the number of participants who had seen misinformation or admitted to posting it unintentionally. These numbers indicate how many participants were qualified to answer behavioral questions about what they do after seeing or posting misinformation.

Table 3.3: Level of misinformation exposure among survey participants.

Survey Question	**Yes**	**No**
Have you ever seen misinformation posted or distributed on social media?	93.3% (942)	6.7% (68)
Have you ever posted or linked to something you later realized was misinformation?	25.7% (260)	74.3% (750)

Registered Analyses

All pre-registered analyses and statistical tests were performed and are described in the following subsections. Tables 3.4-3.6 summarizes the registered analyses for each of the hypothesis tests by research question. For detailed descriptions of the hypotheses and analyses, refer to the Design Table in Appendix C. The “Sample Size” column lists the number of participants whose responses qualified for each paired hypothesis test. The “Interpretation” column follows the standard classification scheme for Bayes Factors [175].

RQ1: How do people respond and think others should respond when they see misinformation on social media?

First, we summarize the interventions that participants report using. Figure 3.1 shows the total number of participants who responded at least once with each possible intervention from Table 3.1. We see that ignoring the misinformation was the most common response at every closeness level. Higher-effort actions (like privately messaging the user, commenting, or creating another post) received relatively more traction when responding to misinformation against close contacts compared with somewhat or not close contacts.

Next, we tested our first hypothesis (**H1.1**), which is that participants would be hypocritical when comparing their opinions about what people should do when they see misinformation with their actual actions when seeing misinformation. For **H1.1**, we found overwhelming evidence (Bayes Factor $>$ 100) that participants believe individuals should expend more effort responding to misinformation on social media than those individuals report actually doing when they

Table 3.4: Pre-registered Bayesian paired hypothesis test results and interpretation for RQ1. Values in italics are for the hypotheses being run with the effort level summed as a robustness check. The p-value for the generalized McNemar's Chi-square Test is also included.

Hypothesis	**Sample Size**	**Null Interval**	**Mean Diff (S.D.)**	**Effect Size**	**Bayes Factor**	**95% HDI for Effect Size**	**Interpretation**	**McNemar Chi-Sq Test**
H1.1: Close contacts	148	$(-\infty,0)$	0.47 (0.88)	0.53	>100	[0.35,0.70]	Extreme evidence for H1	p = 1e-7
			0.99 (1.93)	*0.51*	*>100*	*[0.33,0.67]*	*Interpretation is the same*	
H1.1: Somewhat close	370	$(-\infty,0)$	0.65 (0.92)	0.71	>100	[0.59,0.82]	Extreme evidence for H1	p<2e-16
			1.34 (1.94)	*0.58*	*>100*	*[0.47,0.69]*	*Interpretation is the same*	
H1.1: Not close	880	$(-\infty,0)$	0.45 (0.87)	0.51	>100	[0.44,0.58]	Extreme evidence for H1	p<2e-16
			0.95 (1.87)	*0.51*	*>100*	*[0.44,0.58]*	*Interpretation is the same*	
H1.2: Close v. somewhat	122	$(-\infty,0)$	0.21 (0.84)	0.26	>100	[0.07,0.43]	Extreme evidence for H1	p = 1e-4
			0.16 (1.56)	*0.11*	*6.94*	*[-.07,0.28]*	*Moderate evidence for H1*	
H:1.2 Somewhat v. not	327	$(-\infty,0)$	-.052 (0.88)	-.059	0.170	[-.17,0.05]	Inconclusive at 1/10 threshold	p = 2e-4
			-0.21 (1.52)	*-0.14*	*<1/100*	*[-.24,-.03]*	*Extreme evidence for H0*	
H1.3: Close v. somewhat	1010	$(-\infty,0)$	0.10 (0.58)	0.18	>100	[0.12,0.24]	Extreme evidence for H1	p = 3e-8
			0.16 (1.34)	*0.12*	*>100*	*[0.06,0.18]*	*Interpretation is the same*	
H1.3: Somewhat v. not	1010	$(-\infty,0)$	0.42 (0.81)	0.51	>100	[0.45,0.58]	Extreme evidence for H1	p<2e-16
			0.34 (1.78)	*0.19*	*>100*	*[0.13,0.25]*	*Interpretation is the same*	

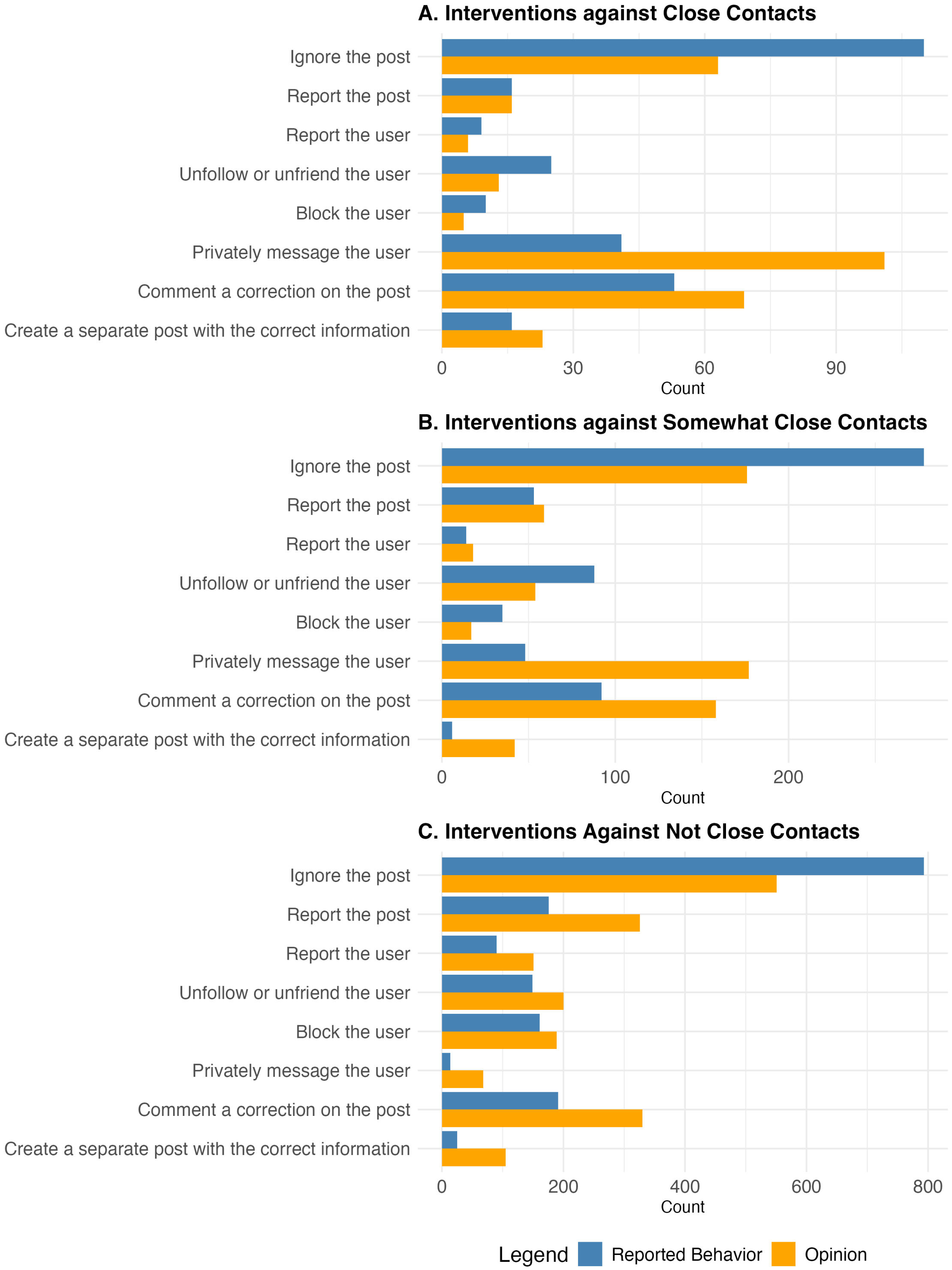


Figure 3.1: Total number of participants who selected each intervention type from Table 3.1 at least once in their behavioral and opinion responses. For comparison purposes, only participants who had seen misinformation at that closeness level have their opinion counts included.

encounter misinformation. This extremely significant result held no matter how close the participant claimed to be to the poster of the misinformation (close contacts, somewhat close contacts, and not close contacts). The effect size was greater than 0.5 in all three closeness cases, indicating a moderate effect size. The unrestricted 95% highest posterior density interval for the effect sizes were [0.35-0.70], [0.59-0.82], and [0.44-0.58] for close, somewhat, and not close contacts, respectively. This result also held the same strength of evidence when the tests were run using a summed effort level rather than a maximum effort level. Figure 3.2A-C shows the distribution of the maximum effort level reported per closeness level.

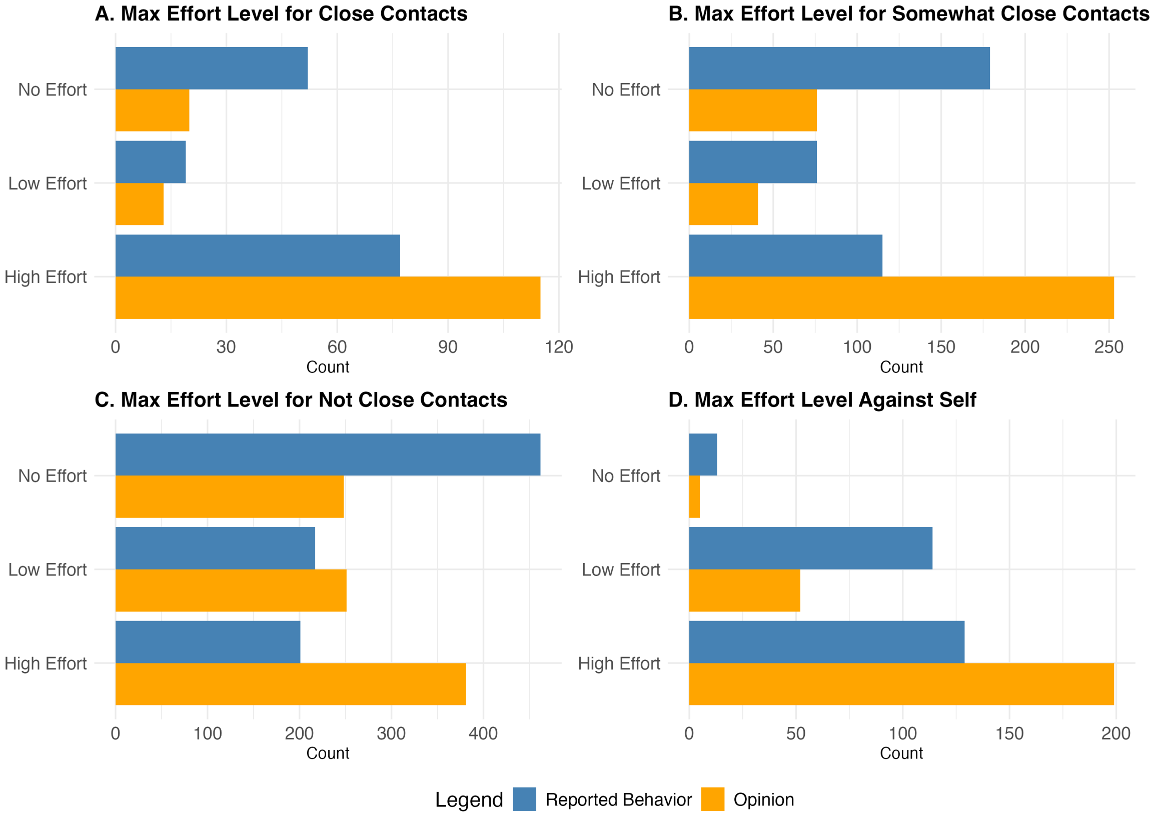


Figure 3.2: Number of participants who reported expending a maximum of no, low, or high effort when seeing misinformation compared with the total number of those same participants who ***believe*** one should expend no, low, or high effort when seeing misinformation at each of the three closeness levels and against oneself.

For **H1.2**, we found that participants responded with more effort when the misinformation poster was a close contact vs. a somewhat close contact (BF $>$ 100) but that there was little difference in responses for somewhat close contacts compared with not close contacts (BF inconclusive). However, for **H1.3**, we found that participants believe that more effort should be expended on close contacts compared with somewhat close contacts and somewhat close contacts compared with not close contacts (BF $>$ 100). Despite the belief that more effort should be put into responding to misinformation posted by a somewhat close contact relative to a not close contact, it seems that participants treated their somewhat close contacts and not close contacts with similar effort levels in practice.

RQ2: How do people respond and think others should respond when they realize they have posted misinformation?

Over 25% of the participants (see Table 3.3) admitted to accidentally posting misinformation at least once. Figure 3.3 summarizes the interventions people claimed to have taken after realizing their mistake. The most frequently reported behavior was deleting the post, followed by updating the post with accurate information.

Table 3.5: Pre-registered Bayesian paired hypothesis test results and interpretation for RQ2. Values in italics are for the hypotheses being run with the effort level summed as a robustness check. The p-value for the generalized McNemar's Chi-square Test is also included.

Hypothesis	**Sample Size**	**Null Interval**	**Mean Diff (S.D.)**	**Effect Size**	**Bayes Factor**	**95% HDI for Effect Size**	**Interpretation**	**McNemar Chi-Sq Test**
H2.1 Self	256	$(-\infty, 0)$	0.30 (0.55)	0.55	>100	[0.41,0.68]	Extreme evidence for H1	p = 1e-13
			1.28 (1.94)	*0.66*	*>100*	*[0.52,0.79]*	*Interpretation is the same*	

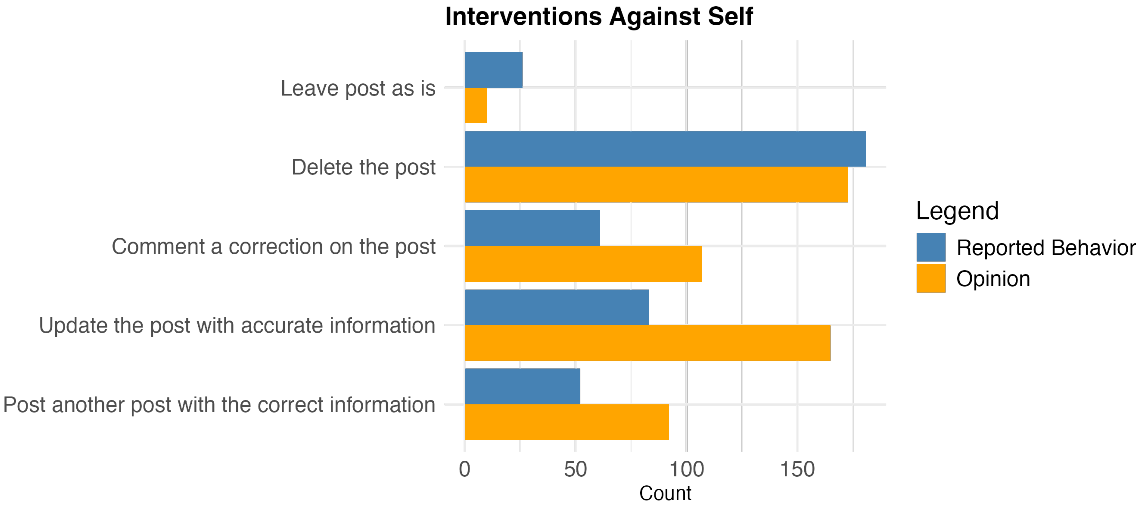


Figure 3.3: Total number of participants who had posted misinformation and selected each intervention type from Table 3.2 at least once in their behavioral and opinion responses. For comparison purposes, only participants who admitted to posting misinformation had their opinion counts included.

We found extreme evidence (BF > 100) that people believe that they should expend more effort to respond to the misinformation they posted compared to what they actually do after realizing they have posted misinformation (**H2.1**). The effect size was 0.55, with the 95% HDI of the effect size being [0.41-0.68]. This result held with the same strength of evidence when the test was run using a summed effort level rather than a maximum effort level. See Figure 3.2D for the distribution of maximum effort level after one has posted misinformation.

RQ3: How do people's responses and beliefs about how others should respond after seeing misinformation differ from their responses and beliefs when they realize they have posted misinformation?

We next investigated how participants' responses and beliefs differ when seeing misinformation versus posting it oneself. To answer this research question, we ran non-directional Bayesian hypothesis tests where we set the null interval to be between [-0.2, 0.2]. For **H3.1**, we found strong evidence that participants respond with more effort when they post misinformation compared with when they see it posted by somewhat close (BF = 79.5) and not close contacts (BF > 100); however, inconclusive evidence that there was a difference in their responses when compared to close contacts.

For **H3.2**, we found extreme evidence (BF > 100) that participants believe that people should respond with more effort when they post misinformation compared to seeing it by not close contacts and inconclusive evidence (BF = 6.92) when comparing their misinformation posts to those posted by somewhat close contacts. Finally, we found very strong evidence (BF < 1/100) towards the null hypothesis that there is not a difference in the level of effort people believe one should use after posting misinformation oneself vs. seeing a close contact post it. These results indicate that participants believe the most effort should be afforded to counter misinformation posted by close contacts or themselves compared with countering misinformation posted by somewhat and not close contacts.

Robustness Tests

We ran three robustness checks. First, we ran the registered hypothesis tests using summed effort values instead of maximum effort values. For H3.1 and H3.2, low-effort actions are excluded from this analysis because there are an unequal number of them described in Tables 3.1 and 3.2. In almost all cases, these tests yielded the same or a similar strength of evidence for the hypotheses. In the few instances where the results diverged, the registered test was inconclusive at the Bayes Factor threshold of 1/10 or 10, while the summed version of the test surpassed the threshold in the same direction (**H1.2**: somewhat vs. not close) or vice versa (**H3.1**: oneself vs. somewhat, **H3.2** oneself vs. close). In only one instance did the interpretation completely differ: for **H3.1** (oneself vs. max of all closeness), the registered test showed strong evidence for H1, whereas the summed test showed strong evidence for H0. Notably, the calculated effect size was positive in both cases. However, in the summed version of the hypothesis test, most of the 95% unrestricted highest density interval (HDI) lay below 0.2, placing it in the null interval.

Second, we used the generalized McNemar's Chi-square Test of Independence for categorical paired data to verify that the interpretation is similar if the data are analyzed categorically rather than as interval data. For every hypothesis, the chi-square test produced a p-value of <= 0.01. This outcome diverged from some of the pre-registered Bayesian analyses, which had found some inconclusive results or evidence pointing towards the null hypothesis for some hypotheses. These discrepancies occurred in tests where the effect size was small, and the HDI overlapped

Table 3.6: Pre-registered Bayesian paired hypothesis test results and interpretation for RQ3. Values in italics are for the hypotheses being run with the effort level summed as a robustness check. The p-value for the generalized McNemar's Chi-square Test is also included.

Hypothesis	**Sample Size**	**Null Interval**	**Mean Diff (S.D.)**	**Effect Size**	**Bayes Factor**	**95% HDI for Effect Size**	**Interpretation**	**McNemar Chi-Sq Test**
H3.1: Self v. close	49	(-0.2,0.2)	0.27 (0.86)	0.31	0.594	[0.010,0.57]	Inconclusive	p = 1e-3
			0.37 (2.26)	*0.16*	*0.125*	*[-.12,0.43]*	*Interpretation is the same*	
H3.1: Self v. somewhat	133	(-0.2,0.2)	0.41 (0.88)	0.46	79.5	[0.27,0.63]	Very strong evidence for H1	p = 9e-9
			0.56 (2.01)	*0.28*	*0.805*	*[0.10,0.44]*	*Inconclusive*	
H3.1: Self v. not close	229	(-0.2,0.2)	0.52 (0.85)	0.61	>100	[0.46,0.74]	Extreme evidence for H1	p<2e-16
			0.67 (1.80)	*0.37*	*31.7*	*[0.23,0.50]*	*Very strong evidence for H1*	
H3.1: Self v. max overall	244	(-0.2,0.2)	0.31 (0.85)	0.36	26.8	[0.23,0.49]	Strong evidence for H1	p = 3e-13
			0.30 (1.86)	*0.16*	*0.070*	*[0.031,0.28]*	*Strong evidence for H0*	
H3.2: Self v. close	1010	(-0.2,0.2)	0.10 (0.75)	0.13	<1/100	[0.071,0.2]	Extreme evidence for H0	p<2e-16
			0.50 (2.15)	*0.23*	*.977*	*[0.17,0.29]*	*Inconclusive*	
H3.2: Self v. somewhat	1010	(-0.2,0.2)	0.20 (0.78)	0.26	6.92	[0.20,0.32]	Inconclusive	p<2e-16
			0.76 (2.14)	*0.35*	*>100*	*[0.29,0.42]*	*Extreme evidence for H1*	
H3.2: Self v. not close	1010	(-0.2,0.2)	0.62 (0.84)	0.74	>100	[0.67,0.81]	Extreme evidence for H1	p<2e-16
			1.71 (2.12)	*0.81*	*>100*	*[0.74,0.88]*	*Interpretation is the same*	
H3.2: Self v. max overall	1010	(-0.2,0.2)	0.031 (0.68)	0.045	<1/100	[-.016,.11]	Extreme evidence for H0	p<2e-16
			0.20 (2.13)	*0.093*	*<1/100*	*[0.031,0.15]*	*Interpretation is the same*	

with the null interval used in that test.

Finally, we ran a categorical analysis to analyze the association between the three variables of interest: maximum effort level used when countering, response type (reported behaviors vs. opinions), and closeness level. The three-way interaction term was not significant, indicating that closeness does not moderate the relationship between effort level and response type. Overall, the results were similar to those in our pre-registered analysis. Effort level interacts with both closeness level and response type, with higher counts of high-effort actions when contacts are closer or when people are asked about their opinions rather than their actual behavior. See Appendix D for additional details.

Exploratory Analysis

RQ4. How do beliefs about responses to misinformation differ based on various demographic factors?

Finally, we examine individual differences in beliefs about how individuals should respond to misinformation. This exploratory analysis complements previous work that examines individual differences in support of misinformation countermeasures implemented by governments, social media companies, and other institutions [255]. For example, several previous studies have found that Democratic individuals are more supportive of platform interventions than Republicans [166, 199, 255], but does this translate into increased support for individual-level measures such as social corrections?

Figure 3.4 shows the maximum effort level participants believe one should exert when encountering misinformation posted by close contacts, somewhat close contacts, not close contacts, or oneself for six demographic variables: age, gender, race, education level, income level, and American political party. See Appendix E for details on the percentage of each demographic category that believes one should respond with no, low, or high effort when encountering misinformation posted by others or oneself. Appendix E also shows the detailed Chi-square test results for each demographic category and closeness level.

For political party affiliations, we find differences in belief in response efforts between partisan groups for close and not close contacts. Strong Republicans supported ignoring posts containing misinformation by close contacts more than any other group, although the absolute difference is $<$10%, which may not have much practical significance. Furthermore, strong Democrats were more likely to support high or low effort responses to not close contacts more than any other group. Notably, we see that, except for not close contacts, at least 70% of respondents in all party affiliations said that one should respond with a high effort action (such as commenting on a correction, updating the post, or messaging the poster).

For age, the chi-squared test shows statistically significant differences in responses among age groups when considering close contacts, somewhat close contacts, and oneself. In general, older participants were more likely to believe one should exert a high level of effort when countering misinformation than younger participants. No significant differences were found between men and women. For racial groups, the only statistically significant difference found was for not close contacts, with Black and Asian Americans more likely to believe in responding with some effort than the other racial groups.

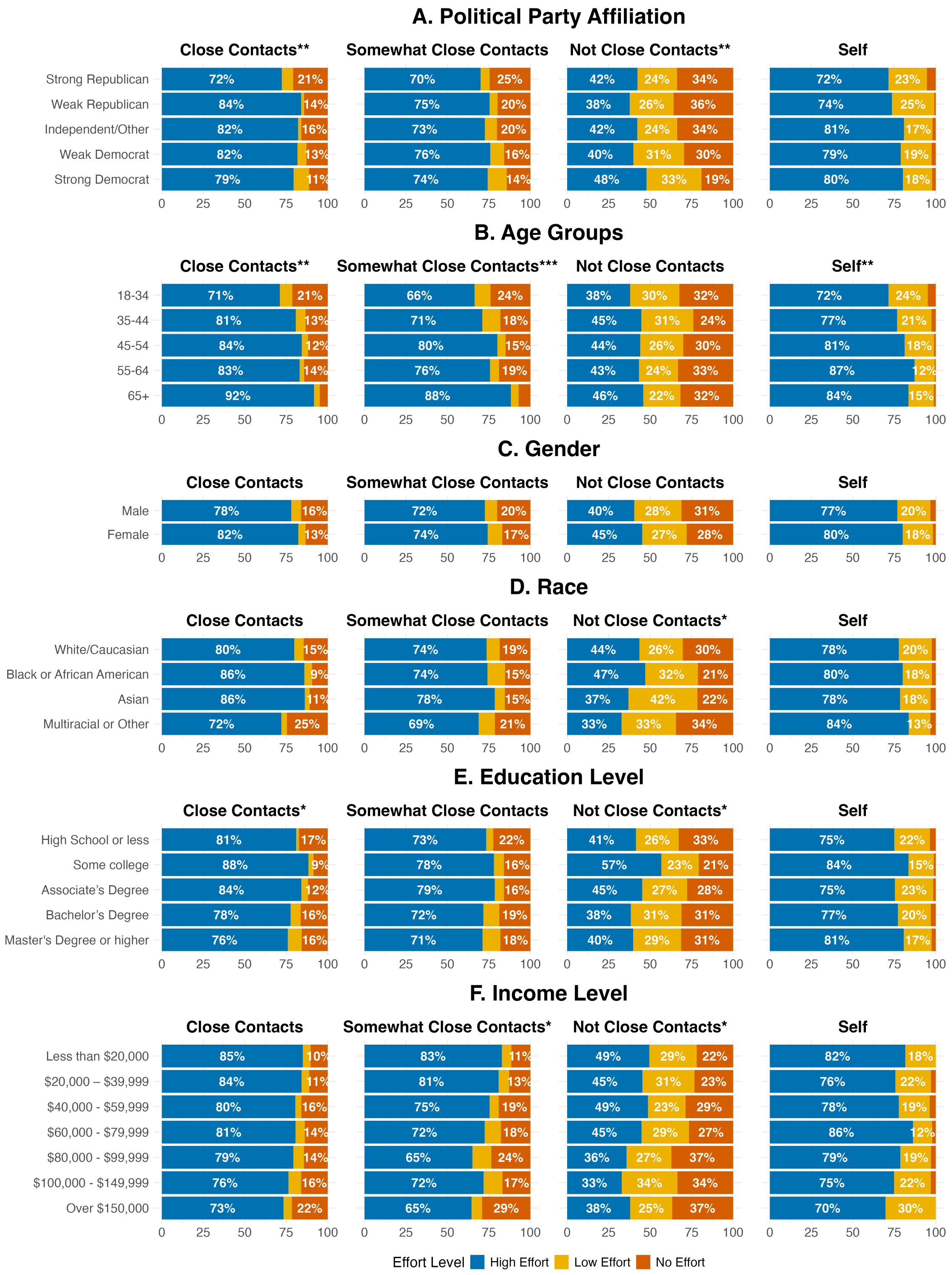


Figure 3.4: Highest effort level participants said one should respond with when seeing misinformation posted by others or oneself broken up by party, age, gender, race, education, and income. Chi-sq tests: $p < 0.05*$, $p < 0.01**$, and $p < 0.001***$.

Finally, the percentage of American residents stating that one should use a high level of effort to counter misinformation drops as education or income level increases. The results from the chi-squared test show statistically significant differences in responses among various education groups regarding close contacts and not close contacts, and among various income groups regarding somewhat close contacts and not close contacts.

3.3.5 Closeness Discussion

In this registered analysis, we compared individuals' beliefs about ideal responses and actual responses to misinformation posted on social media by close contacts, somewhat close contacts, not close contacts, and themselves.

We found overwhelming evidence of hypocrisy in people's responses to misinformation, aligning with our hypotheses (**H1.1, H2.1**). Participants believe others should exert more effort to counter misinformation than they report doing themselves. This pattern holds across all closeness levels, including misinformation posted by oneself, and remains robust to multiple ways of measuring effort. Since there is already a widespread belief that individuals should combat misinformation, efforts to encourage social corrections do not have to convince people to support individual corrections. Instead, they can focus on normalizing these practices and providing strategies to overcome situational constraints (e.g., time and cognitive effort required, social pressures) preventing people from acting.

Furthermore, our results indicate that people not only expect others to exert more effort but also tend to invest more effort themselves when addressing misinformation posted by close contacts compared to those who are somewhat close or not close at all (**H1.2, H1.3**). This increased effort may stem from the impression that correcting a close contact is more likely to be effective due to their relationship, making the effort more worthwhile. Alternatively, people might feel a stronger sense of responsibility to correct a closer contact whose beliefs and behaviors could impact them offline. Additionally, the types of responses differ across closeness as well. For example, people are more likely to privately message a close contact than a less close one. Different approaches may feel more appropriate depending on the source of misinformation. Providing users with a range of options, including private or low-effort methods like reporting, may increase their likelihood of engaging in countering behavior.

When comparing responses to misinformation posted by oneself versus someone else of varying closeness (**H3.1, H3.2**), participants reported putting more effort into responding to misinformation they had posted than to misinformation posted by somewhat or not close contacts. Their beliefs about ideal responses also reflected this pattern. Interestingly, we also found strong evidence that individuals respond with similar levels of effort to misinformation they posted compared with misinformation posted by a close contact. This suggests a similar view of responsibility when the source of misinformation is oneself or a close contact.

Finally, our exploratory analysis revealed demographic differences in beliefs about countering misinformation. Strong Republicans were less likely to believe that high effort should be exerted when countering close contacts, whereas strong Democrats were more inclined to believe some level of effort should be used for not close contacts. This partially aligns with prior research indicating that strong Democrats stood out in their support for institutional countermeasures compared with other partisan groups [255]. We only find this difference holds for not close contacts. Individual-level interventions give people agency to respond to content they believe is misinformation, potentially mitigating distrust in institutional definitions of misinformation. Our findings suggest that this approach to addressing misinformation may be more palatable across the political spectrum.

We found that older Americans were more likely to believe that one should exert high effort to counter misinformation than their younger counterparts. This difference may reflect broader attitudes towards social media, as older individuals are more likely to perceive adverse effects associated with it [24] and, therefore, may be more motivated to address misinformation. Additionally, higher education and income levels were associated with a decreased belief that high effort should be exerted to counter misinformation. Interestingly, higher education and income are also associated with an increased concern and awareness of the negative impact of misinformation [35]. It may be that this concern does not necessarily translate into a belief in the effectiveness or necessity of individual countermeasures. Rather, these concerns may drive greater support for countermeasures on larger scales (e.g., government or platform), which is beyond the scope of this work but should be examined. Additionally, existing literature suggests that higher-income individuals are less generous overall [82, 232], which may extend to efforts to counter misinformation. This preliminary exploratory work can inform future research and platform policies.

3.4 Platform Analysis

Our primary research question here is whether people respond differently to misinformation based on the platform on which it was posted. More specifically:

1. How do people respond when they see misinformation on different platforms?
2. How do people respond when they realize they have posted misinformation on different platforms?

3.4.1 Related Work

Social media users frequently engage with multiple platforms for various purposes, as these platforms can fulfill competing needs [296, 322]. Each platform has different levels of content and account moderation [257], which means users may encounter varying amounts of misinformation depending on the platform [203]. Platforms can also prioritize content in diverse ways, from primarily chronological and time-based information to topic- or location-based arrangements [129]. Given the nature of the interactions, the types of connections, and the content on each platform, users might respond differently to misinformation depending on where it was posted. It is important to consider the types of different platforms and their typical uses.

Platform Categorization

First, we consider the various types of social media platforms. Previous research has categorized the leading platforms in several ways. Kietzmann et al. (2011) classified platforms according to seven primary building blocks: identity, conversations, sharing, presence, relationships, reputation, and groups [155]:

- **Identity** - This block refers to how users disseminate their identity online. Many platforms require users to build profiles, and on many sites, people predominantly retain their real identities (e.g., Facebook, LinkedIn), while on others, they develop virtual identities or use pseudonyms for their usernames (e.g., Reddit).
- **Conversations** - The Conversations block refers to how users communicate with others on the platform. For example, a brief status update on Twitter, which does not require a response from all followers, or a more thoughtful discussion-based post on Reddit.
- **Sharing** - This refers to the transmission of content on the platform and is often related to the platform's purpose. For example, uploading pictures to Instagram or videos to YouTube and TikTok.
- **Presence** - Presence indicates whether other users are online and available to communicate.
- **Relationships** - This building block refers to how users connect with others. These connections may arise from shared interests or topics (e.g., Reddit), from knowing each other in person (e.g., Facebook), or from being fans of others' work (e.g., Twitter, YouTube, TikTok).
- **Reputation** - Reputation refers to how users assess their social standing with respect to others, which can differ based on the platform. Reputation can sometimes be directly evaluated by likes, shares, view counts, comments, and upvotes.
- **Groups** - Finally, the Groups block illustrates how users form communities with each other. Users can create groups like Twitter lists to categorize their friends and followers. Other groups may be open to everyone or require an invitation.

The idea is that platforms tend to focus on three or four blocks at once rather than just one or attempting to address all of them [155]. For example, on LinkedIn, the main building block is identity, but there is also an emphasis on relationships and reputation since it is primarily a career-oriented platform. On YouTube, the main focus is on sharing content, but attention is also given to reputation, groups, and conversations. In the case of Facebook, the predominant block is relationships, but presence, identity, conversations, and reputations are also a part of it [155].

Another frequently cited typology of platforms was developed by Zhu and Chen (2015), and it categorizes platforms into four general categories based on two dimensions: connection type and message type [322]. Connections can either be profile-based, where the focus is on individual users and their profiles, or content-based, where the focus is instead on the content posted. The messages can either be customized for a specific audience or broadcast to the public. Table 3.7 summarizes their four main categories of social media platforms.

While comprehensive, there are aspects that these typologies do not directly incorporate, such as the distinction between private and public content, as well as profiles [296], or the way information is displayed (based on time, location, or topic) [129]. The privacy aspect is likely relevant to countering misinformation, as sharing false information could directly harm one's reputation if associated with one's true identity. I propose a modification to Zhu and Chen's typology that incorporates Kietzmann et al.'s "identity" building block more prominently by dividing each category into two: platforms with predominantly private or anonymous profiles and platforms with predominantly public profiles. Table 3.8 illustrates this updated categorization.

Table 3.7: Social media platform categorization as developed by Zhu and Chen (2015) [322].

	Customized Messages	**Broadcast Messages**
Profile-Based Connections	*Relationship* - Platforms based on users connecting, like Facebook, LinkedIn, WhatsApp	*Self-Media* - Users broadcast to their followers, like Twitter
Content-Based Connections	*Collaboration* - Discussion-driven platforms to find answers or advice, like Reddit or Quora	*Creative Outlets* - Users share their interests and creative pursuits, like YouTube, TikTok, Pinterest

Table 3.8: Modified social media platform categorization.

	Customized Messages		**Broadcast Messages**	
Profile-Based Connections	*Private Relationship:* Facebook, WhatsApp, Messenger, Telegram	*Public Relationship:* LinkedIn, Nextdoor	*Private Self-Media:* Snapchat, BeReal	*Public Self-Media:* Twitter, Threads, BlueSky
Content-Based Connections	*Personal Collaboration:* Reddit	*Public Collaboration:* Quora	*Personal Creative Outlets:* Instagram	*Public Creative Outlets:* YouTube, TikTok, Pinterest, Instagram

In general, users have become more private over time [86], and it is now estimated that nearly half of all social media accounts in the U.S. are set to private [14, 119]. According to a Statista survey in 2018, around 45% of American social media users report that all their social media accounts are private, while another 27% state that some of their accounts are private [88]. Platforms typically do not publicly disclose statistics on the number of private versus public accounts. However, we can make estimates based on various sources, including public opinion polls, the default or typical privacy settings on each platform, and the platform's structure.

When considering the *Relationship* platforms, the sites where most users maintain private or partially private accounts include Facebook, WhatsApp, Messenger, and Telegram. At the same time, LinkedIn and Nextdoor tend to be more public. A 2020 survey conducted for PC Magazine revealed that 57% of Facebook users have accounts that are at least partially private, with nearly 80% saying that they do not share their friends list publicly [119]. Facebook operates as a friendship-based platform rather than a follower-based one, meaning users can usually connect only if both parties agree. For WhatsApp, messages are end-to-end encrypted, and the help center explicitly states that they go to "great lengths to build WhatsApp in a way that helps people communicate privately"[2]. Similarly, Messenger has end-to-end encryption for messages enabled by default[3]. Although Telegram has groups and channels that, in some ways, have similar functionality to a platform like Reddit, it primarily advertises itself as a messaging app that is similar to WhatsApp[4]. Meanwhile, Nextdoor requires users to use their real name and address, as the platform is designed to connect individuals within their local community[5]. By default, users have a public profile on LinkedIn unless they choose to hide it[6].

For the *Self-Media* platforms, Snapchat and BeReal are primarily private [14], while Twitter, Threads, and BlueSky tend to be more public. By default, Snapchat only permits users to be contacted by their "friends"[7]. BeReal is also a friendship-based platform where posts are shared exclusively with friends by default. This platform does not even have a public option: users can share only with at most "friends of friends"[8]. Twitter, BlueSky, and Threads are follower-based platforms. A 2019 Pew Research survey found that only 13% of U.S. adult Twitter users reported keeping their accounts private [242]. Both Threads [276] and BlueSky are considered Twitter competitors, and as of 2023, BlueSky lacked the functionality to support private accounts[9].

When considering the *Collaboration* platforms, Reddit has predominantly anonymous users, while Quora accounts are tied to real identities. Although Reddit is a primarily public platform, according to Reddit, most of their users choose to remain anonymous by selecting screen names[10]. On the other hand, Quora encourages users to use their real names and credentials, and

[2] https://faq.whatsapp.com/595724415641642

[3] https://messengernews.fb.com/2023/12/06/launching-default-end-to-end-encryption-on-messenger/

[4] https://telegram.org/faq#q-what-is-telegram-what-do-i-do-here

[5] https://help.nextdoor.com/s/article/use-your-true-identity

[6] https://www.linkedin.com/help/linkedin/answer/a528138

[7] https://help.snapchat.com/hc/en-us/articles/7012343074580-How-do-I-change-my-privacy-settings-on-Snapchat

[8] https://help.bereal.com/hc/en-us/articles/10444893090205-Audience

[9] https://bsky.social/about/blog/5-19-2023-user-faq

[10] https://support.reddithelp.com/hc/en-us/articles/7420342178324-How-does-being-anonymous-work-on-Reddit

the functionality to answer questions anonymously was removed in 2021[11].

Finally, regarding the *Creative Outlet* platforms, these are generally follower-based instead of friendship-based and are typically public by default unless the user is a minor. YouTube videos are public by default[12], as are Pinterest accounts[13] and TikTok accounts[14], although all of them can be set to private. It remains unclear how many adult users maintain private profiles on these platforms. On TikTok, a Pew Research study found that only about half of adults have ever posted a video, and about 40% had posted a public video [43]. This suggests that most users who post videos do so publicly, but it is unclear how many lurkers have private versus public accounts. Instagram users represent a mix of private and public accounts. A 2020 PC Magazine survey found that about half of Instagram users restrict or filter comments on their posts [119]. Additionally, there is a common practice of creating secondary fake accounts, also known as “finstas”, to preserve privacy and share content with a smaller group of people [14].

Platform Usage

In addition to platform type, it is critical to understand how the different platforms are being used and how platform usage may relate to the issue of countering misinformation. Recent research on platform engagement and advertising has assessed the significance of platform type, indicating that users have varying motivations for using these platforms and that engagement can be highly context-specific [225, 296].

Pelletier et al. (2020) [225] conducted an exploratory survey study and identified four main purposes behind platform usage:

- **Social** - Staying connected with friends and family
- **Informational** - Keeping up with the news and trends
- **Entertainment** - Following sports and celebrities, playing games, and sharing memes.
- **Convenience** - Watching random content when bored or scrolling.

These researchers found that among three major platforms (Facebook, Twitter, and Instagram), users preferred Twitter and Instagram for their social needs, Twitter for their informational needs, and Instagram for their entertainment needs [225]. It is unsurprising that Twitter is preferred for informational purposes, given its classification as a *Self-Media* platform where messages are broadcast to followers. It is similarly unsurprising that Instagram is preferred for entertainment, considering that it is a *Creative Outlet* platform. While Facebook had the largest user base, it also had the lowest actual usage, which may explain why it was not identified as the top platform for social purposes, even though it is a *Relationship*-driven platform.

Similarly, Voorveld et al. (2018) considered 11 dimensions of social media engagement across eight platforms: Facebook, YouTube, Twitter, LinkedIn, Instagram, Pinterest, Snapchat, and Google+ [296]. Besides the four dimensions identified by Pelletier et al., they also considered negative emotions related to the content, negative emotions related to the platform, stimu-

[11] https://productupdates.quora.com/Removing-anonymous-answers

[12] https://support.google.com/youtube/answer/157177

[13] https://help.pinterest.com/en/article/make-your-profile-private

[14] https://support.tiktok.com/en/account-and-privacy/account-privacy-settings/making-your-account-public-or-private

lation, identification, practical use, innovation, and empowerment [296]. Table 3.9 shows which platforms scored the highest and lowest in each dimension, with Google+ excluded as it is no longer operational.

Table 3.9: Social media engagement on various platforms by dimension [296]. The innovaton and empowerment dimensions are excluded, as all platforms scored low.

Dimension	**Highest Engagement**	**Lowest Engagement**
Social interaction	Facebook, Snapchat	YouTube, Pinterest
Informational/Topicality	Twitter, LinkedIn	YouTube, Snapchat
Entertainment	Snapchat, YouTube	LinkedIn, Twitter
Convenience/pastime	Instagram, Facebook	LinkedIn
Negative emotion - content	Twitter, Facebook	Pinterest, LinkedIn
Negative emotion - platform	Facebook	Instagram, YouTube
Stimulation	Pinterest	LinkedIn
Identification	Facebook	LinkedIn, YouTube
Practical use	Pinterest	Snapchat

The researchers found that Facebook and Snapchat, both friendship-based platforms that focus on profile-based connections, ranked highest on the social dimension. Twitter and LinkedIn, primarily public platforms, scored highest on the informational dimension. Snapchat and YouTube, both *Creative Outlet* platforms, scored highest on entertainment. Overall, they noted that the most significant similarity among the platforms was that most of them, except YouTube and Pinterest, were a way to stay informed and up-to-date [296].

Misinformation Exposure

Despite platforms being used for differing purposes, all platforms have misinformation to some extent, even if it is more common on some platforms than others [203, 272]. For example, platforms that prioritize informational content, such as X (formerly Twitter), may have more news-related misinformation than those focused primarily on entertainment, like Instagram and Pinterest. Given that previous research has found that users often feel powerless to combat misinformation because there is so much of it [274], the extent of exposure might be related to the likelihood of countering it.

Finally, previous work in this chapter has demonstrated that closeness influences the willingness to counter, with individuals being more inclined to counter close contacts. Users are more likely to be connected to close contacts on specific platform types, such as the *Private Relationship* platforms like Facebook and WhatsApp. Additionally, any potential social factors at play, such as the desire to maintain interpersonal relationships and concerns about credibility when countering [213], might be less relevant on platforms where users are primarily anonymous, like Reddit. Understanding how different platform types relate to misinformation exposure and countering is a crucial aspect that could help inform future platform policies and public messaging regarding these issues.

3.4.2 Results

RQ1: Responses to seeing misinformation by platform

We are interested in examining any differences in exposure to misinformation and responses among the various platforms. Our subquestions are as follows:

1. Which platforms do most people report encountering misinformation? Which platforms do people report seeing the **most** misinformation?
2. Is the frequency of platform usage connected to exposure to misinformation?
3. How do the actions individuals take when encountering misinformation differ across platforms?

Figure 3.5 shows the platforms where people report seeing misinformation compared to their usage of those platforms. The platforms are sorted by the total number of participants who visit each platform at least weekly, from highest to lowest. The most visited platforms among participants were YouTube, Facebook, Reddit, X, and Instagram, with more than half of the participants claiming they visit these platforms at least once a week.

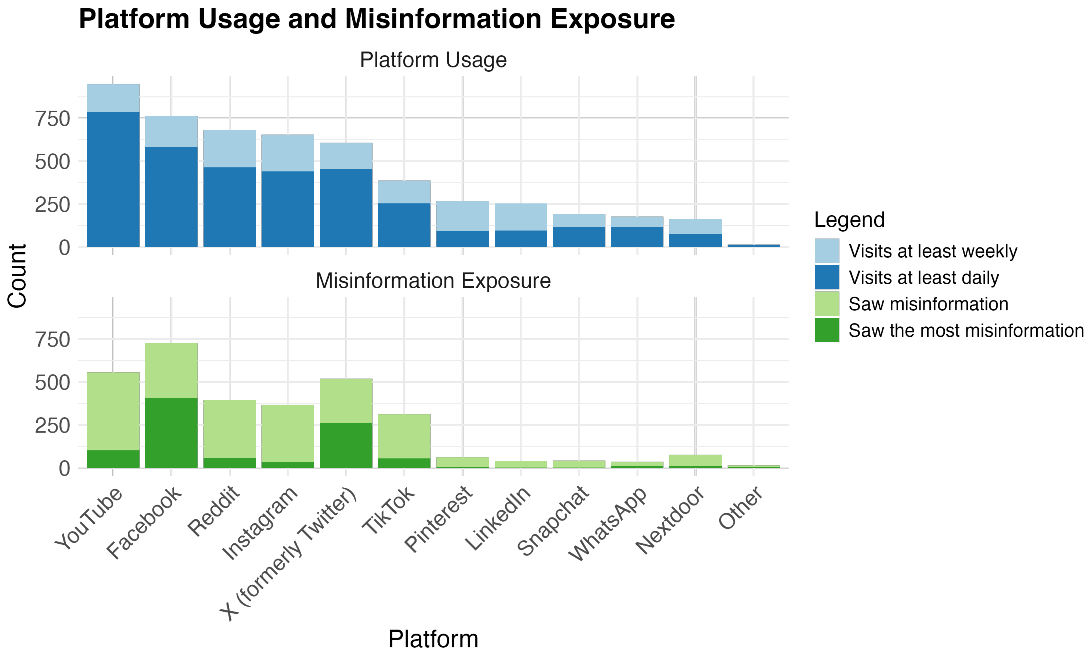


Figure 3.5: The number of participants who use each platform at least weekly or daily compared with the number of participants who have seen misinformation on that platform.

The most frequently used platforms are not fully aligned with those where participants reported encountering misinformation. The total number of participants who claim to have seen misinformation at least once is highest for Facebook, followed by YouTube, X, Reddit, Instagram, and TikTok. In addition, participants were asked to select just one platform on which they believe they have seen the most misinformation, and Facebook (43%) and X (27.8%) dominated.

Table 3.10 displays the number of participants who saw misinformation on each platform, categorized by their closeness to the misinformation poster. The percentages indicate the proportion of people who observed misinformation at each closeness level, calculated from the total number of individuals who saw misinformation on that platform. Since individuals can see misinformation from multiple closeness levels, the percentages in each row total more than 100%.

Table 3.10: Number and percent of people who saw misinformation from close, somewhat, and not close contacts per platform. The top three platforms by percentage per closeness level are bolded. Platforms are sorted by the number of people who have seen misinformation on the platform.

	Contact Type			**Misinformation Exposure**	
Platform	**Close**	**Somewhat**	**Not Close**	**Any**	**Most**
Facebook	**128 (17.6%)**	**305 (42.0%)**	513 (70.7%)	726	**405 (43.0%)**
YouTube	7 (1.3%)	24 (4.3%)	**533 (96.0%)**	555	**102 (10.8%)**
X	10 (1.9%)	57 (11.0%)	484 (93.1%)	520	**262 (27.8%)**
Reddit	1 (0.3%)	13 (3.3%)	**383 (97.0%)**	395	58 (6.2%)
Instagram	22 (6.0%)	85 (23.2%)	310 (84.5%)	367	34 (3.6%)
TikTok	6 (1.9%)	18 (5.8%)	**296 (95.2%)**	311	54 (5.7%)
Nextdoor	1 (1.3%)	6 (7.9%)	69 (90.8%)	76	9 (1.0%)
WhatsApp	**8 (22.2%)**	**13 (36.1%)**	17 (47.2%)	36	9 (1.0%)
LinkedIn	1 (2.5%)	12 (30.0%)	31 (77.5%)	40	1 (0.1%)
Snapchat	**5 (11.6%)**	**11 (25.6%)**	32 (74.4%)	43	2 (0.2%)
Pinterest	2 (3.2%)	4 (6.5%)	57 (91.9%)	62	3 (0.3%)
Other	1 (6.7%)	3 (20.0%)	12 (80.0%)	15	3 (0.3%)
Total	148	370	880	942	942

Participants were more likely to encounter misinformation from close or somewhat close contacts on platforms like Facebook, WhatsApp, Snapchat, and Instagram compared to other platforms. On other platforms, such as YouTube, X, Reddit, TikTok, Nextdoor, and Pinterest, the vast majority of misinformation was posted by contacts who were not close. These differences likely relate to the purpose and functionality of the platforms, with people more likely to follow and communicate with closer contacts on more private platforms like Facebook and WhatsApp compared to other platforms.

Next, we examine the maximum effort level employed by participants when encountering misinformation on the top six platforms with the most misinformation: Facebook, YouTube, X, Reddit, Instagram, and TikTok. Figure 3.6 illustrates that on Facebook and Reddit, participants were more likely to report engaging in high-effort actions to counter misinformation. For Facebook, this difference may be due to the greater amounts of misinformation originating from close and somewhat close contacts on that platform compared to others. For Reddit, it could be because it is a *Collaboration* platform designed for discussion, where most users remain anonymous or adopt a virtual identity, reducing the likelihood of in-person conflicts.

However, if we consider the maximum effort level against only not close contacts among

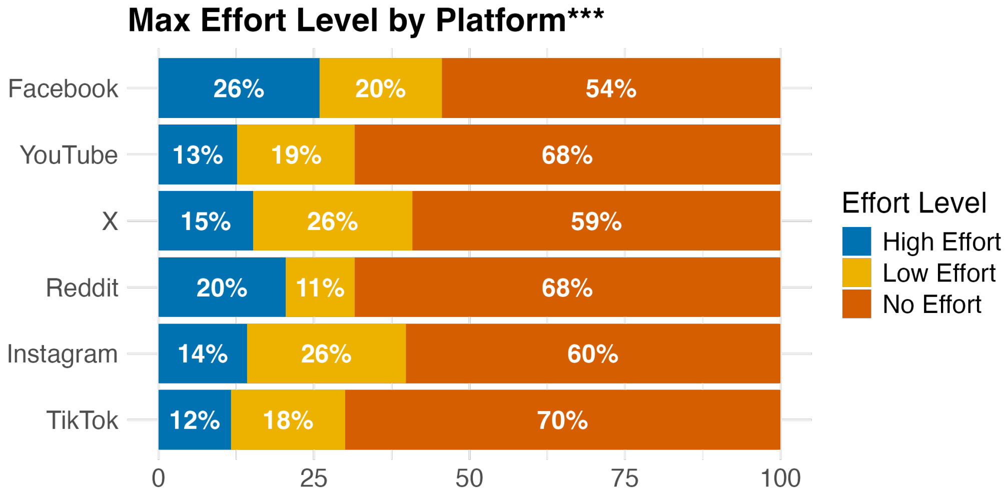


Figure 3.6: Highest effort level participants said they responded with when seeing misinformation posted by others on the top 6 platforms for misinformation. Chi-sq test: $p < 0.001$***.

these top six platforms, differences remain (Figure 3.7). Reddit had the highest percentage of participants who reported engaging in high-effort responses. Participants were more likely to exert effort to counter posts on Facebook, X, and Instagram than on YouTube and TikTok. These differences may be due to the ease of reporting or other platform functionalities and are worth investigating further. Qualitative responses on how closeness and platform may affect countering efforts are further examined in Chapter 4.

RQ2: Responses to posting misinformation by platform

Next, we examine whether there are differences in the posting of misinformation across various platforms. More specifically:

1. Which platforms do people report posting misinformation?
2. How does what people do when they realize they have posted misinformation differ between platforms?

Figure 3.8 shows the platforms where individuals report posting misinformation, whether intentionally or unintentionally. The platforms are sorted by the total number of participants who said that they had unintentionally posted misinformation on that platform, from highest to lowest. Notably, Facebook emerged as the platform with the most posted misinformation, followed by X, and then Reddit and Instagram. Interestingly, the most used platform by participants, YouTube, does not appear in the top three. Table 3.11 shows the top six platforms and details the selected interventions, along with the number of individuals who reported posting misinformation on that platform either intentionally or unintentionally.

Next, we examine the maximum level of effort participants exerted after realizing they posted misinformation on the top four platforms. In this case, we considered only the top four platforms instead of the top six, as no other platform had more than 25 participants admitting to posting misinformation. Figure 3.9 illustrates the maximum level of effort exerted after accidentally

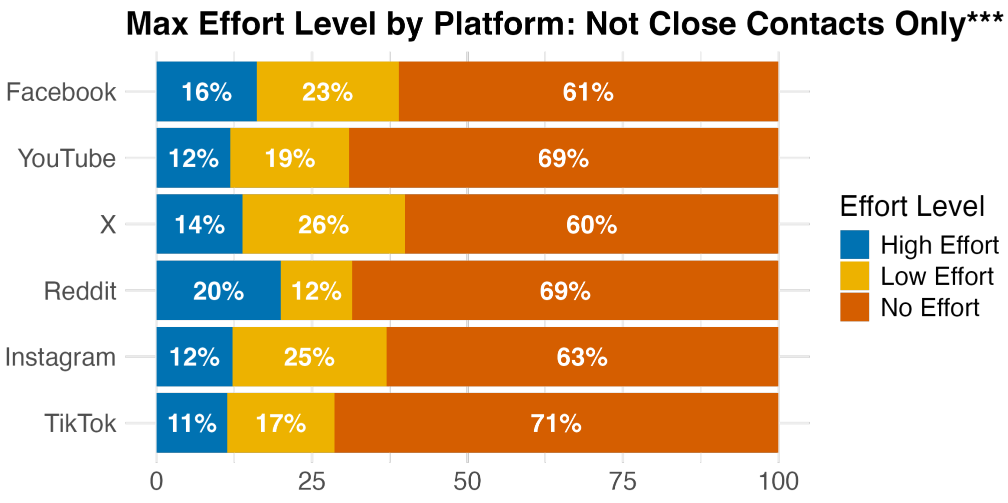


Figure 3.7: Highest effort level participants said they responded with when seeing misinformation posted by **not close contacts only** on the top 6 platforms for misinformation. Chi-sq test: $p < 0.001$***.

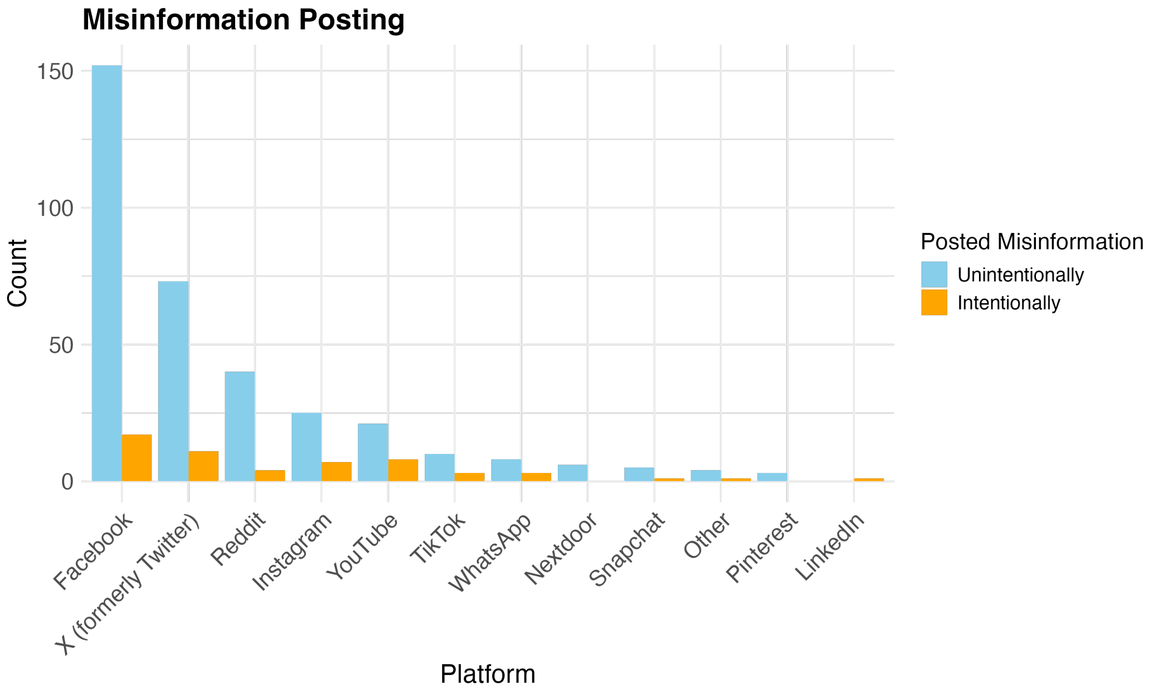


Figure 3.8: Number of participants who admitted to posting misinformation on each platform.

Table 3.11: Summary of interventions selected after posting misinformation.

	Intervention Type					Posted Misinformation	
Platform	**Left post**	**Delete post**	**Comment correction**	**Update post**	**Post correction**	**On Accident**	**On Purpose**
Facebook	13	111	34	48	27	152	17
X	7	57	12	15	20	73	11
Reddit	7	19	9	19	4	40	4
Instagram	3	18	7	6	3	25	7
YouTube	3	9	7	8	3	21	8
TikTok	3	2	3	1	2	10	3

posting misinformation on Facebook, X, Reddit, and Instagram. The chi-squared test of independence was not statistically significant. While there may be differences among platforms, the sample size for most platforms was small, indicating that more research in this area is needed to gain further insights.

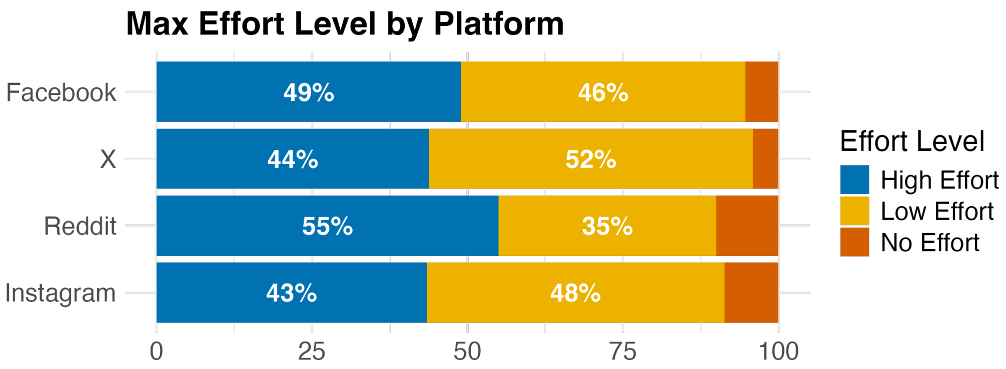


Figure 3.9: Highest effort level participants said they responded with after realizing they had posted misinformation accidentally among the top 4 platforms. Chi-sq test was not significant.

3.4.3 Platform Discussion

Overall, we observed differences in misinformation responses depending on the platform where the content was posted. First, we found that the social media platform is related to misinformation exposure. Participants reported encountering more misinformation on Facebook and X (the second and fifth most used platforms) than on other sites. While misinformation exposure is related to platform usage, Facebook and particularly X overperformed in the survey question "Which platform do you see the most misinformation on?" compared to the number of participants using those platforms.

Furthermore, we found that the maximum effort level in countering misinformation varied across platforms and that proximity to the misinformation poster cannot fully explain these differences. While participants reported seeing more misinformation from close and somewhat close contacts on certain platforms (like Facebook and WhatsApp) than on others (such as YouTube, Reddit, or TikTok), when considering responses regarding only not close contacts, we still observed a greater level of effort on some platforms than others. Facebook, X, and Instagram had the highest percentage of participants saying they engaged in any level of effort to counter misinformation. In contrast, Reddit had the highest percentage of participants reporting that they exerted a high level of effort. These differences may be related to platform social norms, the ease of reporting misinformation, or platform functionalities. For example, Reddit is a more discussion-oriented platform that may allow for higher-effort actions like detailed debunking.

Finally, we found that most participants who admitted to accidentally posting misinformation reported that they did so on Facebook. The second-highest platform was X, with only half as many participants claiming to have unintentionally shared misinformation. Again, this may be related to the functionality and purpose of the platforms. People may be more inclined to post on platforms like Facebook and X while being more passive viewers of content on the *Creative Outlet* platforms like YouTube and TikTok.

3.5 Overall Discussion

This work has several practical implications for promoting public participation in countering online misinformation in educational and technological contexts. First, our research demonstrates the widespread approval of social corrections online, indicating the social desirability of these behaviors. Prior work shows that highlighting the social desirability of reporting misinformation as an injunctive social norm can motivate reporting [110] and decrease the sharing of misinformation [145]. More broadly, there is strong evidence that social signals (e.g., engagement metrics and comments) influence responses to misinformation [26, 76]. However, they can also encourage harmful behavior, such as sharing misinformation that aligns with one's pre-existing beliefs [172]. Therefore, platforms and organizations can successfully promote individual interventions by emphasizing their popularity among users, but they must be careful to avoid inadvertently empowering users with questionable motives.

Additionally, the observed disparity between reported behaviors and beliefs could be leveraged to encourage greater public participation. Research on hypocrisy suggests that one of the most effective strategies for driving behavioral changes is to have individuals publicly commit to pro-social actions, such as signing a pledge, and then be privately reminded of times they have failed to follow through [270]. Public call-outs are less effective, as they may prompt people to save face or rationalize their failures by reducing their support for the targeted behavior [270]. Social media platforms could encourage users, such as those who sign up to contribute to Community Notes programs, to publicly support social corrections or similar measures during educational sessions and regularly remind them of their commitment going forward.

Moreover, the result that people believe more should be done to respond to misinformation than they report doing themselves, in conjunction with the differences we see between platforms, indicates that there may be barriers to employing social corrections or reporting features that could be mitigated by platform design, such as improving transparency, usability, and technical support of these features [321]. Platforms should educate users about their reporting systems when joining and provide periodic updates to keep them informed. Users are also more likely to use reporting features if they perceive them as effective. Therefore, sharing information about the outcomes of reports filed or community notes written can incentivize people to use these programs [310, 321]. Furthermore, pop-up windows have been used on several platforms to ask users if they wish to share content they have not reviewed [74], and this method could be used in scenarios where users delete content. Instead of deleting potentially misleading posts, users could be encouraged to edit or update their posts with accurate information.

Recognizing that susceptibility to misinformation varies across demographic groups [207], educational efforts could tailor strategies to effectively reach different populations [54]. For example, older adults are particularly vulnerable to political misinformation, potentially due to lower digital literacy levels [55]. This age group also tends to support higher-effort responses to misinformation encountered online. Therefore, training efforts for older adults might prioritize digital media literacy over encouraging social corrections. People who are less vulnerable to misinformation, on the other hand, can be promptly educated about operationalizing corrections and leveraging specific platform affordances. Platforms could utilize their internal data to identify these users or implement reputation systems, like X's Community Notes program, where users earn "Rating Impact" scores based on the helpfulness of their contributions[15]. Overall, educational efforts should be designed to account for individual differences in both vulnerability to misinformation and perceived responsibility to counter it.

Lastly, there are a myriad of individual differences beyond demographics that influence vulnerability to misinformation and the likelihood to correct it that should be examined further in future work. For example, evidence suggests that those with a tendency toward analytical thinking are less susceptible to misinformation [93, 226]. These are likely the same individuals capable of providing accurate and meaningful corrections, as some level of cognitive effort is necessary for higher-effort responses to misinformation (e.g., commenting on a correction). Platforms can encourage users to think critically by using accuracy prompts or similar measures [93, 231]. In addition, platforms and other institutions can target educational resources towards those with a propensity to engage in critical thinking.

3.6 Conclusions

3.6.1 Limitations

There are several limitations to this work. First, our sample is not demographically representative of all United States residents. We specifically focused on active social media users to better understand current user behavior on social media platforms. While this targeted sample provides relevant information to platforms about how their users act and what they believe, a more demographically representative survey could provide additional information about less active users who can also influence the spread of misinformation. Additionally, while we collected high-level demographic data, we did not investigate the role of more complex individual features, such as analytical reasoning or values, and how they may interact with one's propensity to intervene against misinformation. We leave this to future work.

[15] https://communitynotes.x.com/guide/en/contributing/writing-and-rating-impact

Next, participants were asked to recall how they had responded to misinformation they had seen on social media in the past, which they may or may not have encountered recently. This could have led to memory or recall errors. Furthermore, we note the possibility of demand effects or other biases (e.g., social desirability) influencing survey responses. We took care to present the survey to participants as generally about misinformation online without including details that may reveal our expectations. Future work could consider using platform data or conducting a field experiment to observe how people respond to misinformation in real-world contexts. For example, platform data on reporting or social corrections could confirm whether people counter closer contacts more than less close contacts and if these behaviors differ by platform.

Additionally, the results linked to RQ3 in the closeness analysis may have limited generalizability due to the fundamental difference in the potential actions one can employ to correct others compared with correcting oneself. We attempted to enumerate commensurate responses to misinformation, such as commenting a correction on someone else's post or one's own post. However, the low-effort actions are, by nature, not equivalent actions (reporting someone else's post vs. deleting one's post). Additionally, there were more listed low-effort actions for responding to others than responding to oneself. We address this in one of our robustness checks, where we compared summed rather than maximum effort levels and excluded all low-effort responses. However, future studies should consider this inherent limitation when conducting this type of analysis.

Finally, we did not investigate possible differences in behavior, not just the beliefs, of various demographic groups. We additionally had limited platform data on the posting of misinformation because the posting of misinformation by participants was strongly concentrated among just a couple of platforms (Facebook and X). Finally, future work could investigate if behavior or beliefs about how to engage on social media are related in any way to support for platform or government measures to counter misinformation. We will explore how misinformation exposure may be related to support of various countermeasures in Chapter 5.

3.6.2 Contributions

This study makes an important contribution to the literature on individual-level interventions against misinformation. Our results indicate that facilitating individual responses to misinformation seen or accidentally posted on social media is a viable approach to reducing the spread of misinformation and even preventing belief in it. Using a large sample of active social media users in the US, we demonstrate the widespread belief that individuals should counter misinformation despite a tendency to not always act on this belief themselves. The nature of responses and the willingness to expend effort vary based on the user's relationship with the misinformation poster and the platform, highlighting opportunities to educate the broader population about different ways to take action depending on their perceived situational constraints. These insights inform efforts to encourage public participation in mitigating the impact of misinformation and suggest ways that platforms can enhance their countering tools to empower users to engage more actively in maintaining the integrity of their online information environment.

Chapter 4

Improving User-based Countermeasures

In Chapter 3, I analyzed current user opinions and behaviors concerning user-led countermeasures. This chapter focuses on determining whether the gap between the intention to engage in user-based measures and actual behavior can be narrowed by utilizing media literacy training efforts. Can media literacy serve as a means to empower individuals and combat misinformation at the user level? If so, educational initiatives and training games can be developed by the government, platforms, civic society, or even individuals. They are scalable and, therefore, practical.

The primary research question for this chapter is: How can we improve the usage of user-based countermeasures? More specifically,

1. Are media literacy and training games effective at **improving misinformation detection?**
2. Are media literacy and training games effective at **improving the countering of misinformation?**
3. What factors, such as the **poster or platform**, are associated with one's willingness to counter misinformation?

4.1 Introduction

As misinformation continues to affect societies around the world, much recent research has focused on developing and deploying effective interventions [46, 79, 131, 167]. One of the most studied intervention types involves media literacy and digital literacy as a preventive measure [79]. Media literacy encompasses many different types of interventions, ranging from short tips [121], in-person training sessions [193], to fake news games [36, 176].

Most media literacy experiments focus on improving participants' skills to better discern truth from falsehoods and improve their critical understanding of the media they encounter [141]. Although many studies also investigate the effectiveness of media literacy in behavioral outcomes, the behaviors studied typically focus on reducing the frequency of harmful, risky, or antisocial behaviors such as engaging with or sharing misinformation or participating in risky sexual encounters. To our knowledge, no studies have yet focused on improving the willingness and ability of participants to counter it [141]. We seek to fill this research gap by running an experiment focused on increasing motivated individuals' willingness, knowledge, and ability to counter misinformation.

This study was conducted with 23 motivated government analysts, where participants were shown a series of social media posts and asked if they believed the posts to be true or false. Participants were also shown several explicitly false or misleading social media posts and asked to describe if and how they would respond. A survey was administered before and after an interactive, in-person training session to determine if misinformation detection ability improved and if countering willingness increased. We also included open-ended questions that allowed participants to explain their reasoning so that we could qualitatively analyze the factors associated with willingness to engage in countering behavior.

This chapter will be broken into two main results sections: **improving misinformation detection** and **improving misinformation countering**. The results of this case study will help inform how to encourage and improve user-based countermeasures, such as social corrections, thus contributing to broader efforts to combat misinformation.

4.2 Related Work

4.2.1 Media Literacy Interventions

Media literacy interventions consist of educational initiatives designed to enhance the public's civic discourse by improving critical thinking ability when reading media content [121, 141]. One type is the development and usage of fake news games. These games include the Bad News Game [36], Go Viral! [200], Troll Spotter [176], and Harmony Square [249]. They are designed to be an interactive and fun way to help players detect misinformation [186, 200].

A related concept to media literacy is the theory of inoculation, sometimes referred to as "prebunking." Inoculation includes interventions such as preemptive warning messages or other anticipatory interventions meant to "inoculate" people, much like a vaccine would, from later believing that misinformation or harmful content [180]. Similarly, media literacy is also intended to improve participant resilience when encountering misleading, harmful, or false messages.

The effectiveness of media literacy as a preventative measure against misinformation has been widely debated in the literature. There is a lack of consensus on whether it is effective, which types are effective, how effective it is, how long the effectiveness lasts, and in which contexts it is most effective [27, 36, 141, 200].

4.2.2 User-Based Interventions

In the context of countering misinformation, user-based interventions refer to actions that social media users can take when directly engaging with misinformation [28, 142]. For example, social media platforms typically allow users to report other users or posts [214]. Social media users can also employ social corrections. Social corrections attempt to debunk the misinformation poster by publicly commenting on their post, privately messaging the user, or other related means [28].

User-based measures are an essential type of misinformation countermeasure. Although most platforms employ some automated moderation, they also rely on social media users to report anything those algorithms miss. In addition, users can comment and engage in social corrections to help debunk misinformation. Having a trusted messenger, such as a friend or family member, debunk misinformation has been found to be effective in several studies [28, 48, 182, 275, 303]. User-based interventions are a vital tool in the fight against misinformation, and there is currently little to no work being done utilizing media literacy interventions to improve those measures.

4.3 Data and Methods

The participants in this study were government analysts who had signed up for social cybersecurity training. They participated in a two-week training exercise called "OMEN" (Operational Mastery of the Information Environment) [157], which ran from 02/05/24 to 02/16/24 in Orlando, FL. This study was conducted on 02/07/24 during this training exercise.

4.3.1 OMEN Overview

OMEN is a project administered by the Office of Naval Research designed to teach analysts how to evaluate their online information environment. The first week of OMEN was reserved for various training sessions during which participants learned about social cybersecurity and how to use various software tools. During the second week, analysts participated in the OMEN game, a training game that teaches analysts and decision-makers how to detect and counter misinformation on social media. It 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 that generally matches what the analysts would encounter on their day job. See our associated technical report [157] for more details on the design of the OMEN game, the creation of the storyline, the collection of data, the learning objectives, and the lessons learned.

4.3.2 Participant Demographics

Twenty-three participants completed the study by filling out the pre-test survey, attending the training sessions, and completing the post-test survey. All participants were members of the defense community from one of the "Five Eyes" nations (US, UK, Canada, Australia, and New Zealand). A high-level summary of participant demographics is provided below.

- **Gender**: 19 men (82.6%), 4 women (17.4%)
- **Race**: 20 white (87.0%), 3 mixed or other (13.0%)
- **Age**: The average age was 35.6 years old, with a standard deviation of 10.5 years. The median age was 34, and the range was 21 to 58. The age distribution:
  - 18-24: 3 (13.0%)
  - 25-34: 9 (39.1%)
  - 35-44: 7 (30.4%)
  - 45-64: 4 (17.4%)

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

- **Country**: 16 United States, 3 Canada, 2 Australia, 2 New Zealand
- **Education**: All participants had more than a high school education, with most participants having at least a Bachelor's degree:
    - Some college or Associate's Degree: 8 (34.8%)
    - Bachelor's Degree: 9 (39.1%)
    - Master's Degree: 3 (13.0%)
    - Professional or Doctorate Degree: 3 (13.0%)
- **Political Ideology:** On a Likert scale of 1-5, with "very liberal" as a 1 and "very conservative," the average ideology was fairly moderate at 2.83, and the median was a 3.
    - Very liberal: 2 (8.7%)
    - Liberal: 6 (26.1%)
    - Moderate: 10 (43.5%)
    - Conservative: 4 (17.4%)
    - Very conservative: 1 (4.3%)
- **Social Media Platforms:** Participants were asked which social media platforms they interact with as part of their job and which platforms they use outside of work. The top platforms used for work were X, Facebook, Instagram, YouTube, and LinkedIn. The top platforms for personal use were Facebook, YouTube, Facebook Messenger, Instagram, and Reddit. These results are summarized in Table 4.1.

4.3.3 Survey Design

The assessment survey was implemented in Qualtrics and had two sections: misinformation detection and misinformation countering. The pre and post-tests had identical structures but included different, randomly selected posts.

Part 1: Misinformation Detection

In the first section of the survey, participants were presented with a set of 16 social media posts, randomly selected from a pool of posts from four categories. These categories were national news, misinformation (low-credibility and/or misleading news), local news, and pink slime (imposter local news). The participants saw four posts from each category in a random order and were asked a series of questions for each post. They received the following instructions:

> In this section, you will be shown a series of 16 generic social media posts. If an organization or group posted the content, those posts will show a username and profile picture. Posts from regular users were modified and anonymized.
>
> Any account that has a blue checkmark indicates that at least one mainstream social media organization considers that account to be a news organization / media company. These accounts may vary in accuracy and bias level.
>
> **NOTE**: You may click on links and image link previews.

Table 4.1: The number of participants that used various social media platforms for work and for personal reasons.

Platform	**Work Count**	**Personal Count**
X (formerly Twitter)	20	10
Facebook	18	20
Instagram	15	14
YouTube	13	17
LinkedIn	12	12
TikTok	8	4
Telegram	8	2
Reddit	7	13
Facebook Messenger	5	15
Snapchat	4	9
Tumblr	4	1
Pinterest	3	4
Discord	3	9
WeChat	2	1
Mastodon	2	2
Twitch	2	4
Nextdoor	1	1
Threads	0	1
None	2	1
Other	3	0

The participants were then asked the following questions for each post:

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, Somewhat trustworthy, Neither trustworthy nor untrustworthy, Somewhat untrustworthy, Untrustworhty]*
3. How confident are you in your answers to the previous two questions? *Accuracy: [0-10], Trustworthiness: [0-10]*
4. Do you believe a local reporter wrote this post? *[Yes, No, Unsure]*
5. Would you share this post online (for example, through Facebook or Twitter)? *[Definitely yes, Probably yes, Might or might not, Probably not, Definitely not]*
6. Do you believe the poster of this message is trying to influence you or the audience of this post? *[Definitely yes, Probably yes, Might or might not, Probably not, Definitely not]*
7. Please elaborate on the reasons for your answer to the previous question on influence. *[Free response]*

Several of these questions, particularly those concerning trust and local reporters, are not analyzed in this work and were included in the survey for a concurrent study on "pink slime." Likewise, local news and pink slime posts are also not analyzed in this work. The results of the pink slime analysis can be found in Christine Lepird's dissertation [177].

Part 2: Countering

In the second section, participants were shown four explicitly false posts and then asked what, if anything, they would do if they encountered those posts on their social media feeds. They received the following instructions:

> In this section, you will be shown a series of four false posts and then asked how you would respond.

When presented with these false posts, the participants could select from the list of possible responses shown in Table 4.2. This list of responses was developed for Chapter 3 [158] but was extended to include an "Other" option. Responses labeled as "Low Effort" refer to indirect or quick actions, while responses labeled as "High Effort" apply to actions that directly engage with the misinformation content and require more time. Participants who selected "Other" were prompted to write in their response, and the level of effort for their response was manually categorized as no, low, or high.

More specifically, the participants were asked the following questions for each post:

1. If you saw this post on your social media feed, what would you do? (select any that apply) *[Options described in Table 4.2]*
2. *[If a "High Effort" action was selected]* What would you write in your comment or post? *[Free response]*
3. Do you think your answer would change depending on how well you knew the person or organization posting it? *[Definitely yes, Probably yes, Might or might not, Probably not,*

Table 4.2: Actions one can take when engaging with misinformation on social media.

Response	**Effort Level**
Ignore the post	No Effort
Report the post	Low Effort
Report the user	Low Effort
Block the user	Low Effort
Unfollow or unfriend the user	Low Effort
Privately message the user	High Effort
Comment a correction on the post	High Effort
Create a separate post with the correct information	High Effort
Other	-

Definitely not]

4. Please elaborate on why or why not your response would change depending on the person or organization posting it. *[Free response]*
5. Do you think your answer would change depending on which platform you saw this post on? *[Definitely yes, Probably yes, Might or might not, Probably not, Definitely not]*
6. Please elaborate on why or why not your response would change depending on which platform you saw this post on. *[Free response]*

Post-Test Only Questions

The only difference between the pre-test and post-test Qualtrics surveys is that the post-test survey included four summary questions to assess the effectiveness of the training sessions. The participants were given the following instructions and questions:

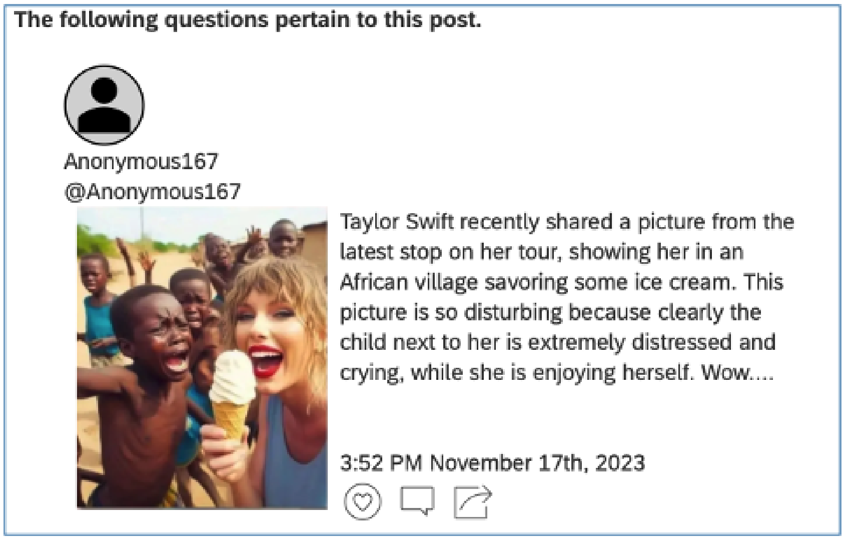
> Instructions: Please rate your level of agreement to each of the following statements
> using the below rating scale

1. This training has helped me become better at recognizing pink slime news. *[Strongly disagree, Somewhat disagree, Neither agree nor disagree, Somewhat agree, Strongly agree]*
2. This training has helped me become better at recognizing misinformation. *[Strongly disagree, Somewhat disagree, Neither agree nor disagree, Somewhat agree, Strongly agree]*
3. This training has helped me become more knowledgeable about how to counter misinformation. *[Strongly disagree, Somewhat disagree, Neither agree nor disagree, Somewhat agree, Strongly agree]*
4. Which of the following techniques that were taught in training did you utilize when answering the questions in this survey? Select all that apply.
   - Source: Clicking on the link and reading the article
   - Source: Checking a news website's "About" page
   - Author: Looking up the author(s) of the article

- Other Sources: Reading upstream - clicking links/sources in the article
- Other Sources: Lateral reading - searching keywords or searching for similar stories
- 3rd Party: Looking up the bias/accuracy ratings of the news organization in question
- Experts: Checking Fact-Checking sites
- Other (please describe) *[Write-in]*

Post Selection

We designed the survey posts to resemble typical social media posts, complete with profile pictures, user names, time stamps, and the appearance of buttons to like, comment, and share. Figure 4.1 shows an example post from the countering section. The full set of posts used in this study can be found in Appendix F.



The following questions pertain to this post.

Anonymous167
@Anonymous167

Taylor Swift recently shared a picture from the latest stop on her tour, showing her in an African village savoring some ice cream. This picture is so disturbing because clearly the child next to her is extremely distressed and crying, while she is enjoying herself. Wow....

3:52 PM November 17th, 2023

Figure 4.1: An example of a false post shown to participants.

The topics in the posts were selected to be apolitical and timely. They included health (COVID-19, vaccines), science (climate change, flat earth theories), and current entertainment topics (the Barbie movie, Taylor Swift). To account for possible differences in the difficulty of assessing each post, we had a group of experts review the posts for both difficulty and quality. Between four and six CASOS PhD students were tasked with reviewing each post. The reviewers completed the same survey described above but with additional review questions.

For the posts in the misinformation detection section, the reviewers were asked the following questions:

1. How easy or difficult do you think assessing the accuracy and trustworthiness of this post would be for an average American social media user? *[Extremely easy, somewhat easy, Neither easy nor difficult, Somewhat difficult, Extremely difficult]*
2. Should we include this post in the survey? *[Yes, No]*

3. Any comments on this post? *[Free response]*

For the posts in the countering section that only included explicitly false or extremely misleading posts, the reviewers were asked the following questions:

1. Would the average American government employee think this post was true or false? *[True, False, Unsure]*
2. How easy or difficult do you think it is to counter this post for the average American government employee? *[Extremely easy, somewhat easy, Neither easy nor difficult, Somewhat difficult, Extremely difficult]*
3. Should we include this post in the survey? *[Yes, No]*
4. Any comments on this post? *[Free response]*

Based on experts' comments, some posts were removed if they were deemed poor examples or confusing. In the misinformation detection section, two out of the 19 misinformation posts tested and four out of the 16 actual news posts tested were removed. The two misinformation posts were removed because the reviewers noted that it required watching the videos to fact-check, which would not have been conducive in the testing environment. Among the four actual news posts that were removed, one was removed because the headline changed and no longer matched the content of the linked article, one because the study linked in the news article moved and therefore was difficult to verify, one because while the news organization was reputable, it failed to note that the study they quoted had a conflict of interest, and one because most reviewers had trouble verifying the legitimacy of the source. None of the 12 false posts in the countering section were recommended for removal.

The remaining posts were sorted by post type (real/national news, misinformation, local news, pink slime, and false posts for the countering section) and by average difficulty score (rated on a 1-5 Likert scale). They were then randomly divided into the pre- and post-tests. This method was employed to control the item-wise difficulty of the posts and to maintain an equal number of posts per post category in the pre- and post-test. The difficulty scores for the relevant questions in this study are summarized below:

- **Misinformation Posts** - Average difficulty scores ranged from 1.5 to 3, with an overall average of 2.33
- **Real/National News Posts** - Average difficulty scores ranged from 1.6 to 3, with an overall average of 2.31
- **Countering Posts** - Average difficulty scores ranged from 1.2 to 3, with an overall average of 2.09.

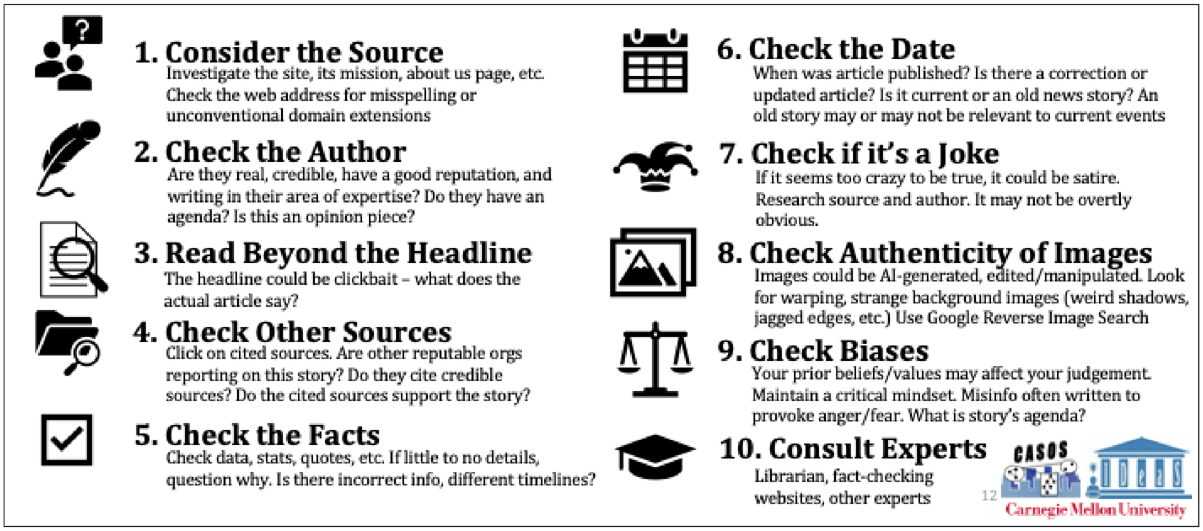
The posts in each category were sorted by their average difficulty score and divided into quartiles. Four misinformation posts, four national news posts, and four countering posts, one from each quartile of difficulty, were randomly selected for the pre-test. The same method was applied to randomly select the posts for the post-test. All participants were shown the same posts in the pre-test and post-test because there was a concern that they would discuss among themselves during the breaks throughout the day.

4.3.4 Training Sessions

After the pre-test survey was administered, participants received training in three main areas: misinformation detection, pink slime detection, and countering misinformation. The pink slime training was an interactive 30-minute session administered by Christine Sowa Lepird [177] and is not covered further in this work. I led the other two training sessions. Following the training and a break for lunch, participants took the post-test. By the end of the day, participants were briefed on what they did well and the most frequently missed questions.

Misinformation Detection Training

Participants were given a 45-minute training adapted from my work from previous chapters, various university library trainings, and information from FactCheck.org [154, 233]. First, participants were provided with background information on the topic, including formal definitions of misinformation and disinformation. In addition, participants were given a summary of the various characteristics of misinformation, covering eight main types along with their purpose, audience, and news context (see Section 1.3). Next, they were given ten steps to consider when assessing the accuracy of online information, followed by several examples illustrating each of the ten steps. An example was shown for each major type of misinformation. Figure 4.2 shows the training slide summarizing all ten steps. Participants had access to the slides and could use this information as a reference. The training was interactive, with respondents describing how they would approach the examples presented in the slide deck.



1. Consider the Source
Investigate the site, its mission, about us page, etc. Check the web address for misspelling or unconventional domain extensions

2. Check the Author
Are they real, credible, have a good reputation, and writing in their area of expertise? Do they have an agenda? Is this an opinion piece?

3. Read Beyond the Headline
The headline could be clickbait – what does the actual article say?

4. Check Other Sources
Click on cited sources. Are other reputable orgs reporting on this story? Do they cite credible sources? Do the cited sources support the story?

5. Check the Facts
Check data, stats, quotes, etc. If little to no details, question why. Is there incorrect info, different timelines?

6. Check the Date
When was article published? Is there a correction or updated article? Is it current or an old news story? An old story may or may not be relevant to current events

7. Check if it's a Joke
If it seems too crazy to be true, it could be satire. Research source and author. It may not be overtly obvious.

8. Check Authenticity of Images
Images could be AI-generated, edited/manipulated. Look for warping, strange background images (weird shadows, jagged edges, etc.) Use Google Reverse Image Search

9. Check Biases
Your prior beliefs/values may affect your judgement. Maintain a critical mindset. Misinfo often written to provoke anger/fear. What is story's agenda?

10. Consult Experts
Librarian, fact-checking websites, other experts

Figure 4.2: Summary training slide on misinformation detection.

Countering Misinformation Training

Participants were given a 30-minute interactive training session adapted from my literature review work on why and how to counter misinformation effectively. The training was broken down into three main parts:

1. **Why People Should Counter Misinformation** - We discussed the most common reasons why people do not counter misinformation and how those concerns can be addressed [267, 274]. The main reasons discussed were:
    - *Level of Impact* - The first concern discussed was the potential impact one could have when countering misinformation. Research showing the effectiveness of individual-level debunking, especially from trusted messengers, was shared with participants [182].
    - *Time Restraints* - Countering misinformation can be time-consuming [267], so participants were encouraged to concentrate on those most likely to change their minds or on individuals closest to them.
    - *Conflict Avoidance* - Maintaining harmony in relationships is another primary concern [213]. Participants discussed respectful ways to listen and debunk, and were guided on indirect actions they could take instead, such as reporting, blocking, unfollowing, or writing a community note.
    - *Credibility Concerns* - Some individuals raised concerns about countering misinformation on topics where they are not experts. Participants were encouraged to refer to experts and fact-checking organizations, or to engage in indirect actions.
2. **Common Logical Fallacies** - We reviewed logical fallacies and how to spot them. Examples and explanations were adapted from several university research guides[2][3][4]. The discussion primarily focused on the following fallacies:
    - *Slippery slope* - This fallacy occurs when someone believes that a specific event or action will lead to a series of increasingly severe events without any evidence. It appeals to one's emotions by exploiting the fear of this possible chain of consequences without any real evidence to support this claim.
    - *Straw man* - The straw man fallacy occurs when someone misrepresents their opponent's argument by fabricating or exaggerating their position, making it easier to debate them. This tactic is often done to make one appear more reasonable in comparison.
    - *Ad hominem* - Ad hominem attacks involve targeting an opponent's personal characteristics to undermine their position. While this argument has legitimate uses if one's characteristics are relevant, it is frequently used to insult and degrade individuals without engaging with their actual arguments.
    - *Bandwagon* - This fallacy occurs when individuals argue that something must be true or good simply because it is popular or widely supported by the public. While arguments that have popular support may indeed be true, the fallacy lies in assuming that they must be true solely due to their popularity.
    - *False causation* - False causation is often employed by those who intentionally spread misinformation. It refers to the idea that if one event occurs before another, the first

[2] https://libguides.princeton.edu/c.php?g=982190&p=7102155

[3] https://owl.purdue.edu/owl/general_writing/academic_writing/

[4] https://writingcenter.unc.edu/tips-and-tools/fallacies/

event must have caused the second. However, correlation does not always imply causation.

3. **Effective Interventions** - Finally, participants were trained on the types of interventions and debunking efforts that are effective [182, 282]. At the individual level, they were encouraged to use trusted community messengers when debunking misinformation. They were also encouraged to ask questions to engage the critical thinking skills of those who believe the misinformation and to demonstrate empathy to emphasize that they share the same overall goals and concerns. Participants were provided with high-level examples of the following types of debunking:

    - *Fact or logic-based debunking* - The suggested method was the use of a "fact" sandwich, or starting and ending with the accurate information while debunking the myth or fallacy in the middle to better emphasize the facts over the misinformation [182].
    - *Empathy-based debunking* - Empathy-based debunking may sometimes be effective in countering conspiracy theories by pointing out the similarities between the targeted or scapegoated group and one's own group. [41, 179, 219]
    - *Source-based debunking* - This method aims to reduce the credibility of those spreading misinformation or conspiracy theories. Ridicule or humor has been shown to be effective for general audiences, but it may not be advisable when countering the beliefs of an individual [179, 219]

    At the organizational level, they were encouraged to focus on careful and transparent research dissemination to prevent the public or news organizations from misinterpreting their results. They were also encouraged to utilize social media campaigns when appropriate to promote counter-narratives against pervasive falsehoods that often go unchecked on social media, such as anti-COVID-19 vaccination narratives [282].

4.3.5 Analysis Methods

Part 1: Misinformation Detection

Table 4.3 summarizes the measures used in Part 1 of this study. Descriptive statistics will be summarized for all measures. Paired t-tests will be used to compare the average accuracy and confidence scores from the pre-test and post-test.

Part 2: Countering

Table 4.4 summarizes the measures used in Part 2 of this study. The chi-squared test will be used to compare the pre- and post-test effort level measures. Paired t-tests will be used to compare the average poster and platform effects per participant in the pre- and post-tests.

Textual analysis will be used to analyze the open-ended questions associated with the poster and platform effect measures. To analyze the free-text responses, we used code mapping to identify common themes throughout the responses. Code mapping, sometimes referred to as affinity diagramming, is frequently used when analyzing open-ended survey data [140, 254]. We first reviewed all the responses for each question and then sorted related comments into groups

Table 4.3: The measured variables from Part 1 of the surveys.

Variable(s)	Definition	Values
Misinformation Accuracy National News Accuracy	Response to "What do you believe is the accuracy of the content in this post?"	1 = True 2 = Somewhat true 3 = Neither true nor false 4 = Somewhat false 5 = False
Misinformation Confidence National News Confidence	Response to "How confident are you in your answers to the accuracy question?"	0–10 where: 0 = Very unsure 10 = Very confident

using Apple's Freeform program[5], which is a digital whiteboard. Whenever a participant noted something like "same reason as my previous answer," we placed those in the same group as their previous related response. We iteratively developed the codes by reviewing the responses a second time. Then, we labeled each group of responses with various factors and merged related factors into overarching themes. The primary themes are discussed in the Results section.

4.3.6 Ethics Information

The Carnegie Mellon University Institutional Review Board (IRB) approved this study, numbered "STUDY2023_00000429." They determined the study was exempted from full review because it involved a "benign behavioral intervention." All participants were randomly assigned a user ID to link their pre- and post-training results, but their names were not collected. The survey collected informed consent from all participants. Participants were not paid by the study but instead were paid their typical government salary.

4.4 Results: Misinformation Detection

4.4.1 Accuracy

For the four misinformation and four national news posts in both the pre- and post-tests, participants were asked to rate what they believed the accuracy of those posts was on a 1-5 Likert scale. Table 4.5 displays the average and median number of questions participants answered correctly.

[5] https://www.apple.com/newsroom/2022/12/apple-launches-freeform-a-powerful-new-app-designed-for-creative-collaboration/

Table 4.4: The measured variables from Part 2 of the surveys.

Variable(s)	Definition	Values
Effort Level	Response to “If you saw this post on your social media feed, what would you do?”	**Set to the max value:** Ignore post = 0 Report post = 1 Report user = 1 Block user = 1 Unfollow/unfriend user = 1 Message user = 2 Comment correction = 2 Create post = 2 Other = [write-in]
Poster Effect	Response to “Do you think your answer would change depending on how well you knew the person or organization posting it?	1 = Definitely yes 2 = Probably yes 3 = Might or might not 4 = Probably not 5 = Definitely not
Platform Effect	Response to “Do you think your answer would change depending on which platform you saw this post on?”	1 = Definitely yes 2 = Probably yes 3 = Might or might not 4 = Probably not 5 = Definitely not

An accuracy score of 3, 4, or 5 for misinformation posts was labeled as correct, while an accuracy score of 1 or 2 for actual news posts was labeled as correct. The median values are shown to account for potential item effects.

Overall, participant detection of misinformation remained relatively unchanged, as it started high. The average number of misinformation posts correctly identified increased only slightly from the pre-test to the post-test, from 87% to 88% correct. Unsurprisingly, the two-sided paired t-test shown in Table 4.6 comparing the average accuracy scores was not significant. Participants were likely more familiar with misinformation detection techniques than we expected, considering the nature of their professions.

Table 4.5: The average and median number of questions answered correctly by participants.

	Pre-Test		**Post-Test**	
Questions	Average	Median	Average	Median
Misinfo. Posts	3.48 (87%)	4 (100%)	3.52 (88%)	4 (100%)
Real News Posts	2.43 (61%)	3 (75%)	3.30 (83%)	3 (75%)

Table 4.6: Two-sided paired t-test results (df = 22) and the estimated effect size, Cohen's *d*, for the average accuracy scores given by each participant as defined in Table 4.3.

	Pre-Test	**Post-Test**	**Paired T-Test**		
Questions	Average (SD)	Average (SD)	*t*-value	p-value	Cohen's *d*
Misinfo. Posts	4.21 (0.63)	4.20 (0.45)	0.065	0.95	0.0135
Real News Posts	2.33 (0.70)	1.77 (0.58)	3.23	0.0038	0.675

However, the detection of real news did increase over time, with only 61% of the questions answered correctly on average in the pre-test compared to 83% in the post-test. The average assigned accuracy scores also improved, rising from an average score of 2.33 in the pre-test to 1.77 in the post-test, indicating that participants were more likely to believe that the actual news posts were true in the post-test. The two-sided paired t-test comparing the average accuracy scores was statistically significant.

4.4.2 Confidence

A concern with media literacy training is that it could lower respondents' confidence in their ability to distinguish true from false news. This possible unintended consequence could make people more skeptical of all news, including accurate information, which tends to be substantially more prevalent than false news. We did not observe this in our data. Table 4.7 shows that the overall average confidence score per question type remained relatively unchanged from the pre-test to the post-test. The two-sided paired t-test comparing the average confidence scores in the pre- and post-tests for both the misinformation posts and for the true news posts were not statistically significant.

Table 4.7: Two-sided paired t-test results (df = 22) and the estimated effect size, Cohen's *d*, for the average self-reported confidence scores on a scale 1-10 given by each participant as defined in Table 4.3.

	Pre-Test	**Post-Test**	**Paired T-Test**		
Questions	Average (SD)	Average (SD)	*t*-value	p-value	Cohen's *d*
Misinfo. Posts	8.65 (1.21)	8.22 (2.12)	1.14	0.27	0.24
Real News Posts	8.38 (1.30)	8.67 (1.37)	-1.15	0.26	-0.24

4.5 Results: Countering Misinformation

4.5.1 Selected Interventions

Participants were presented with four clearly marked false posts during both the pre- and post-tests. When asked if and how they would respond to the misinformation shown, the participants could select one or more responses from the options described in Table 4.2. The total number of times participants selected each intervention type is summarized in Figure 4.3.

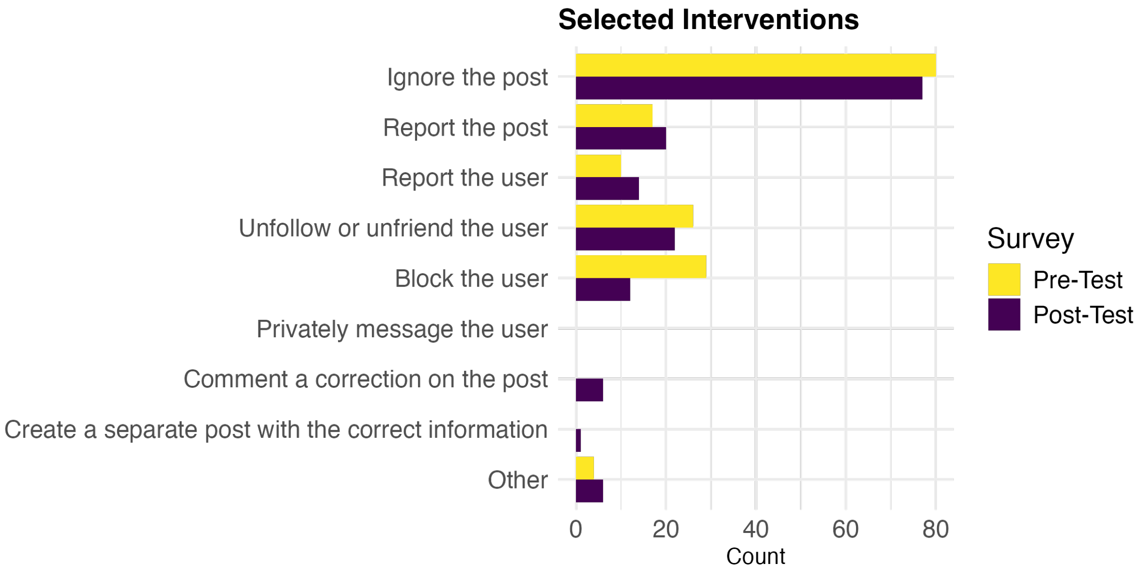


Figure 4.3: The number of times participants said their answer would change depending on either the platform or the poster over both the pre- and post-training surveys.

Unsurprisingly, these results are fairly similar to those observed in Chapter 3 (see Figure 3.1), with ignoring the post being the most frequently selected option. However, many low-effort actions were selected, such as reporting or blocking the user. High-effort actions, such as commenting a correction on the post, were selected more frequently in the post-test.

4.5.2 Effort Level

We found that the maximum amount of effort participants selected to counter any misinformation posts increased from the pre-training survey to the post-training survey (see Table 4.8). More people said they would engage in high-effort actions, with this increase primarily coming from those who were already engaging in low-effort actions. The proportion of participants who stated they would not counter any misinformation posts remained unchanged. The chi-squared test comparing the counts in the effort level categories was significant, with a p-value of 0.049. However, due to the small sample size, the chi-squared approximation may be inaccurate as the expected counts per cell are low.

Table 4.8: The percentage and number of participants whose maximum effort level was in each of the three effort level categories described in Table 4.2.

	Pre-Test	**Post-Test**
No Effort	34.8% (8)	34.8% (8)
Low Effort	65.2% (15)	43.4% (10)
High Effort	0% (0)	21.7% (5)

Table 4.9 shows the contingency table showing the number of participants in each effort level category for both the pre- and post-tests. In the post-test, most participants either maintained the same effort level they selected in the pre-test or increased their maximum effort level. Only three participants (13.0%) opted for a lower effort level. However, the generalized McNemar's chi-squared test, which is applicable for paired data, was not significant with a p-value of 0.16. While the traditional chi-squared test analyzes whether the variables are related, the McNemar's test specifically analyzes whether subjects switched their responses from the pre- to post-tests in a consistent direction. The lack of significance may be related to the small overall sample size.

Table 4.9: The contingency table showing the number of participants that were no, low, and high effort in the pre vs. post-test.

			Post-Test	
		No Effort	**Low Effort**	**High Effort**
	No Effort	5 (21.7%)	2 (8.7%)	1 (4.3%)
Pre-Test	**Low Effort**	3 (13.0%)	8 (34.8%)	4 (17.4%)
	High Effort	0 (0%)	0 (0%)	0 (0%)

4.5.3 Platform and Poster Impact

For each post, participants were asked whether their answers would change depending on the person or organization that posted it or the platform on which they saw the misinformation. Figure 4.4 shows the number of times participants selected each possible answer in the pre- and post-test surveys. "Probably Not" and "Definitely Not" were the most frequently chosen responses for both poster and platform. However, across all the posts shown, 15 participants (65.2%) and seven participants (30.4%) responded with "Probably Yes" or "Definitely Yes" for at least one post when asked if the poster or the platform, respectively, would influence their answer. Participants were more likely to say that the misinformation poster would impact their countering response than the platform.

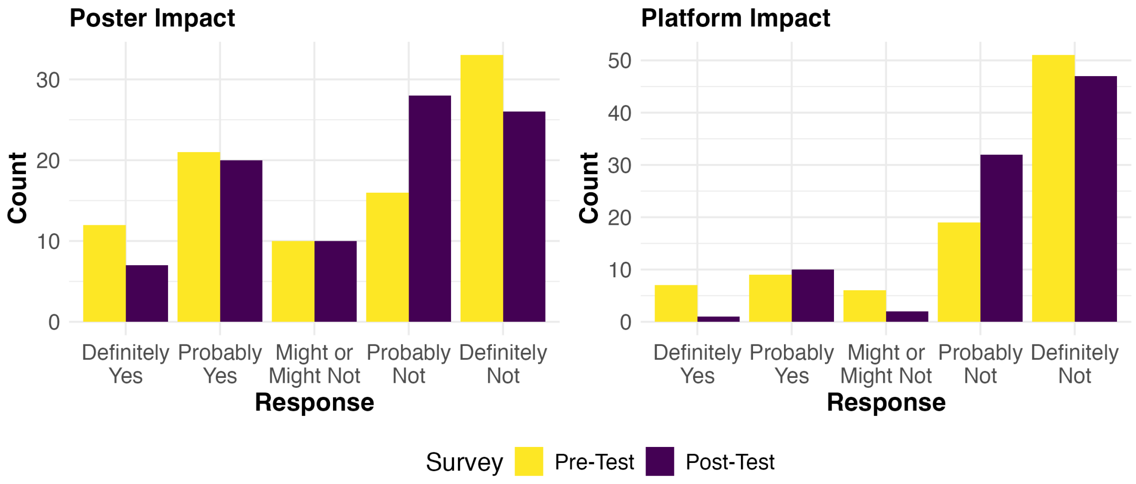


Figure 4.4: The number of times participants said their answer would change depending on either the platform or the poster in both the pre and post-training surveys.

Table 4.10 shows the average scores for the effects of the poster and the platform in both the pre- and post-tests. The two-sided paired t-test, which compared the average scores by participant was not significant for either the poster effect scores or the platform effect scores. Given that the poster and platform effect scores are relatively consistent across the pre- and post-tests, we can combine the total counts for both to better compare the overall poster and platform effects. Figure 4.5 shows the total number of times participants selected each possible answer over both surveys. The chi-squared test comparing the poster with the platform counts was statistically significant, with the p-value < 0.001.

Table 4.10: Two-sided paired t-test results (df = 22) and the estimated effect size, Cohen's *d*, for the average poster and platform effect scores given by each participant as defined in Table 4.4.

	Pre-Test	**Post-Test**	**Paired T-Test**		
Questions	Average (SD)	Average (SD)	*t*-value	p-value	Cohen's *d*
Poster Impact	3.40 (1.11)	3.49 (0.99)	-0.30	0.77	-0.063
Platform Impact	4.07 (1.07)	4.24 (0.80)	-0.95	0.35	-0.20

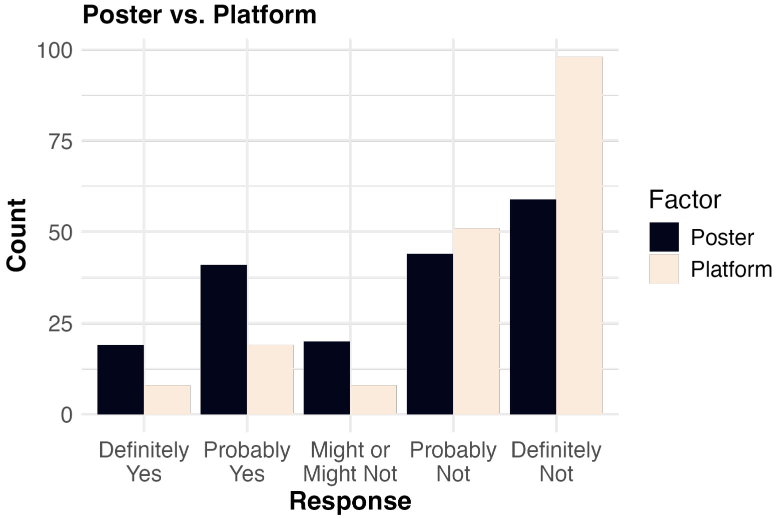


Figure 4.5: The number of times participants said their answer would change depending on either the poster or the platform combined over both the pre- and post-training surveys.

4.5.4 Factors Affecting Countering Actions

The survey asked participants to explain their reasoning about whether and how the misinformation poster or the platform would affect their countering behavior. When elaborating on their responses regarding how these factors would impact their countering efforts, we identified four main recurring themes:

1. *Platform / Account Preferences:* This theme addresses the participant's preference, or lack thereof, regarding which types of posters or platforms they are more likely to engage in countering actions.
2. *Content:* Content refers to how the content of the misinformation post affects an individual's likelihood of engaging in countering efforts.
3. *Platform / Account Features:* This theme focuses on the features of the misinformation accounts and social media platforms and their potential influence on whether someone chooses to take action.
4. *Impact:* Finally, many participants discussed the long-term consequences of their potential countering actions or lack of actions.

All relevant comments fall into these four main themes. Other comments were either irrelevant, supplementary to the participants' main comments, or expressed that they did not care about the post. These other comments are classified under a theme called "Other." Table 4.11 summarizes the major themes and how frequently they were mentioned when considering the misinformation poster or platform, sorted in descending order of total overall mentions. Table 4.12 displays the specific sub-themes mentioned by the participants when considering the misinformation poster. Similarly, Table 4.13 shows the sub-themes when considering platform.

Figure 4.6 illustrates the total number of times each sub-theme was referenced when considering either the misinformation poster or the platform. It emphasizes how prominent the *Platform*

Table 4.11: The number of unique participants and total mentions of each major theme when considering the misinformation poster or the platform.

	Poster		Platform		Overall	
Theme	# Unique	# Total	# Unique	# Total	# Unique	# Total
Platform/Account Preferences	13 (56.5%)	40	13 (56.5%)	55	17 (73.9%)	95
Content	18 (78.3%)	44	12 (52.2%)	20	20 (87.0%)	64
Platform/Account Features	4 (17.4%)	7	3 (13.0%)	10	6 (26.1%)	17
Impact	5 (21.7%)	10	3 (13.0%)	5	7 (30.4%)	15
Other	8 (34.8%)	12	8 (34.8%)	16	12 (52.2%)	28

Table 4.12: The number of unique participants that mentioned each sub-theme and the total number of times each sub-theme was mentioned when participants were asked to consider the poster of the misinformation.

Theme	Sub-Theme	# Unique	# Total
Account Preferences	More effort if closer to poster	13	40
Content	Content is too far gone	13	21
	Content is clearly false / easy to debunk	9	12
	Possible offline harms / very serious	7	9
	Familiarity with or care about the content	3	3
Account Features	More effort for verified / news accounts	3	5
	Less direct effort / more indirect for organizations or less close contacts	2	3
	More effort if recognizable source	2	2
	More effort for accounts with large followings	1	1
Impact	Avoid conflict	1	4
	Perceived lack of countering impact	3	4
	Time-intensive to counter	2	3
Other	Unsure of poster's motive	1	1
	Unrelated comment	7	11

Table 4.13: The number of unique participants that mentioned each sub-theme and the total number of times each sub-theme was mentioned when participants were asked to consider the misinformation platform.

Theme	Sub-Theme	# Unique	# Total
Platform Preferences	Treat all platforms equally	11	47
	Anonymity	1	5
	Platform preference	2	3
Content	Content is too far gone	4	6
	Content is clearly false / easy to debunk	4	9
	Possible offline harms / very serious	5	5
Platform Features	Reporting functionality	1	6
	Platform type	2	2
	Ability to contact poster directly	1	2
Impact	Audience	2	4
	Platform is serious about misinformation	1	1
Other	Don't care	3	5
	Unrelated comment	6	11

/ Account Preferences theme was: closeness to the poster is a huge factor when considering the likelihood of countering. At the same time, most participants indicated that they had no platform preferences.

Platform and Account Preferences

Most participants (73.9%) mentioned preferences in some way across both surveys and factors. When considering the misinformation poster, many participants (13 or 56.5%) repeatedly stated that they prefer engaging with close contacts over less close contacts. These participants suggested that if they knew the misinformation poster, they would be more inclined to participate in direct debunking efforts, whether on social media, via private messages, or in person. This sentiment was by far the most frequently mentioned comment overall when participants were asked if their responses would change depending on the poster. Two examples of quotes from participants are shown below.

> "If I knew the person, I would be more inclined to messaging them about it first."
>
> "If I knew the poster personally i [sic] would privately message them to try and keep them from posting fake news."

On the other hand, while most participants (13 or 56.5%) mentioned the theme of platform preferences, most of them mentioned it because they believe they treat all platforms equally (11 or 47.8%). Only two respondents (8.7%) mentioned having a platform preference. Two examples of quotes are provided below.

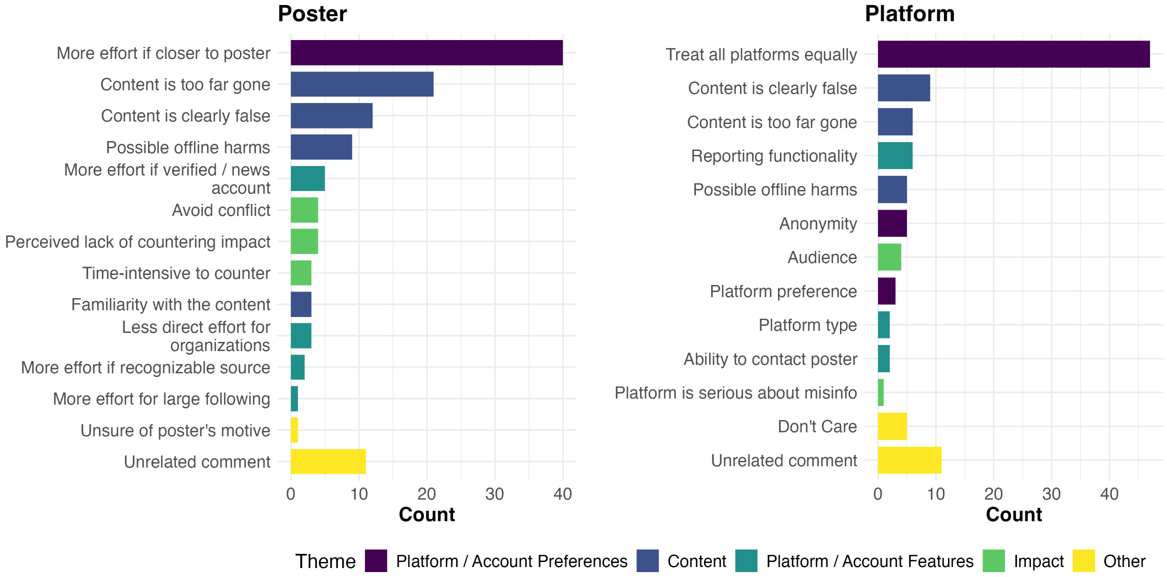


Figure 4.6: This figure displays the total number of times over both surveys that participants mentioned a specific sub-theme when asked if and how their answer would change depending on the misinformation poster and the platform.

> "I am not likely to engage with users on sites where my identity is directly tied to the account. Accounts like Reddit, where I am more anonymous, makes discussion easier to partake in and exit from."
>
> "The platform would not determine my decision."

Content

The content of the post was one of the most frequently cited reasons for engaging or not engaging in countering efforts. Twenty respondents (87%) mentioned content at least once when explaining their reasoning. Specifically, if participants perceived the content as extreme or incredulous, many expressed that they thought any effort would be wasted. For example:

> "I just don't think i [sic] could change the mind of someone who believed the earth was flat."
>
> "If they believe that they would be too far gone for reasoning."

Conversely, if something was easy to debunk, such as a straightforward or factual error, a topic they were knowledgeable about, or if the post had the potential for severe offline consequences, participants indicated they would be more likely to engage.

> "i [sic] would report this one and try to get it to stop promulgating because i think it could be harmful."

Platform and Account Features

Four participants (17.4%) mentioned the significance of account characteristics and said they would be more inclined to intervene if the poster was a verified account, a news agency, a recognizable source, or had a large following.

> "It was a verified channel posting it so the misinformation will spread faster. If it was a random person posting it, I might not repost it and just ignore it instead"
>
> "I would be more likely to report this if it was from a more recognizable source."

While many participants mentioned that they treat all their social media platforms equally, three participants (13%) expressed that they believed they would be more likely to respond on certain platforms than others due to the features of those platforms. For example, one participant stated:

> "If I were in a channel specific to that topic, I may look to validate/invalidate the content. But on a platform like facebook [sic] I would be less likely to give it a second look."

The ease with which one could report a post or contact the misinformation poster was also mentioned as a possible factor. For example:

> "Misinformation is everywhere, trying to fight it is exhausting and not every platform has the same method to report it. "
>
> "On different platforms it might be easier to directly contact the poster."

Impact

Seven participants (30.4%) mentioned the potential impact of countering, and most cited it as a reason for their reluctance to take action. They expressed concerns that debunking would be too time-consuming and have little to no effect, that they lacked sufficient knowledge to counter the post, or that they wanted to avoid conflict.

> "Where do I even begin? An anon account posting disinfo is hardly worth my time, but there are so many misguided people out there, it's easier to report and block rather than engage."
>
> "I feel that reporting would do little to change anything."

One participant mentioned impact positively and said that taking action on some platforms that take moderation seriously may be more impactful than those with less moderation. However, another participant felt differently:

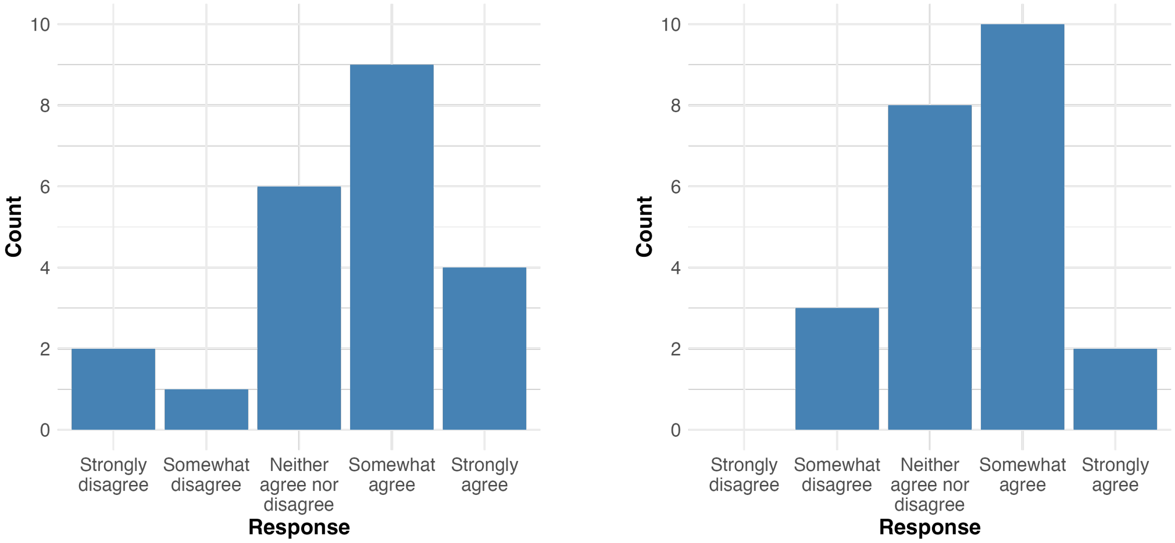
> "I would be more likely to report it on an application or site with worse media literacy."

4.6 Participant Feedback

Feedback from participants after the session was generally positive. At the end of the post-test, participants anonymously answered feedback questions, allowing us to evaluate the effectiveness of the training sessions. First, participants were asked if they felt the training improved their ability to recognize misinformation. They indicated their beliefs on a Likert scale of 1 to 5, with 1 meaning they "strongly disagreed" that the training was helpful and 5 meaning they "strongly agreed." Similarly, they were asked whether they believed the training helped them become more knowledgeable about countering misinformation. Table 4.14 shows the high-level summary statistics, while Figure 4.7 shows the distribution of the feedback scores. In both cases, the median response was a 4 ("somewhat agree"). Considering that many of these analysts were already highly skilled at detecting misinformation, it is not surprising that some did not find the training helpful. However, the majority of participants did. Only three of the 23 participants indicated they disagreed with either feedback statement.

Table 4.14: The average, standard deviation, and median feedback score on a 1-5 Likert scale, where 1 represents "strongly disagree" and 5 represents "strongly agree."

Feedback	**Average (SD)**	**Median**
This training helped me become better at recognizing misinformation	3.55 (1.14)	4
This training helped me become more knowledgeable about how to counter misinformation	3.48 (0.85)	4



(a) Part 1: Misinformation Detection Feedback

(b) Part 2: Countering Ability Feedback

Figure 4.7: This figure summarizes the responses to the feedback questions at the end of the study, indicating whether participants agreed or disagreed that the training was helpful.

Next, we asked participants to select which techniques covered in the training sessions, if any, they used when answering the questions in the post-test survey. Figure 4.8 summarizes the results. The most frequently selected answer was to click on the link and read the article. Most participants employed all the listed techniques except for fact-checking websites. This result demonstrates that the techniques we taught were helpful, although we may want to dedicate more time to expert fact-checking websites in future training sessions.

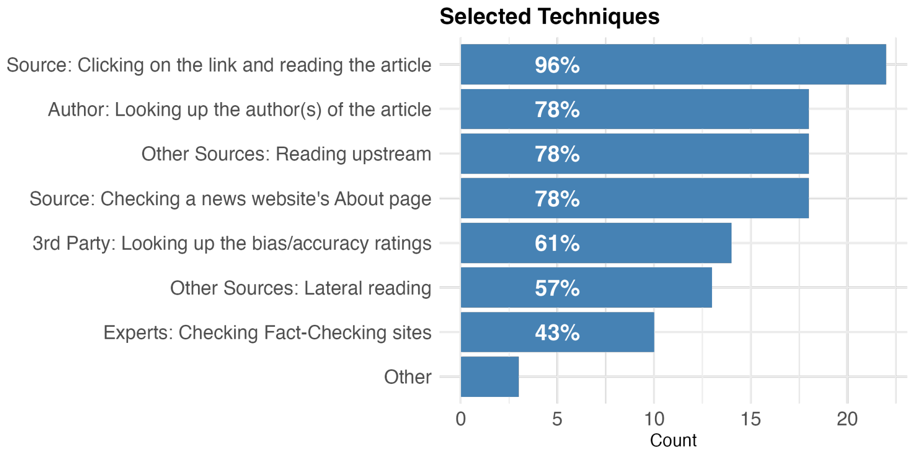


Figure 4.8: The number of participants who said they used each technique in the post-training survey.

4.7 Discussion

We conducted an experiment to assess the effectiveness of media literacy training on improving the detection of misinformation and countering it. The evidence for improved misinformation detection was limited, partly because participants already excelled at identifying misinformation in the pre-test, resulting in minimal improvement in the post-test that was not statistically significant. However, actual news detection increased in the post-test, with more participants accurately classifying real news stories as true (see Table 4.5). Participants' confidence in their ability to differentiate real news posts from misinformation remained steady in the post-test (see Table 4.7), indicating that the training did not unintentionally make participants more skeptical of all news posts.

Next, we analyzed the effectiveness of the training on participants' willingness and likelihood of utilizing countermeasures. We found that more respondents claimed they would intervene with more effort and more directly in the post-test compared to the pre-test (see Table 4.8). However, the number of participants who stated they would exert no effort remained unchanged. This increase in high-effort actions came primarily from individuals who were already engaging in low-effort countering actions, such as reporting or blocking.

We also qualitatively analyzed the participants' explanations regarding how their likelihood of countering would change based on the account posting the misinformation and the platform on which it was posted. Overall, we found that although the **Content** theme was mentioned by the highest number of unique participants (20), **Platform and Account Preferences** were the most frequently cited factors across all posts. As illustrated in Figure 4.6, proximity to the poster and platform neutrality dominate all other factors. These results suggest that individuals who hold these beliefs feel strongly about them and expressed the need to mention them multiple times throughout the survey. Other frequently discussed themes included **Platform and Account Features** and the potential **Impact** (or lack thereof) when countering.

Overall, the feedback from participants was positive. They rated the usefulness of both training sessions with a median Likert score of 4 (“somewhat agree”). Additionally, most participants applied the techniques we covered during the training. These results are encouraging, considering the nature of the participants’ profession and the fact that many were already proficient at identifying most misinformation stories. Each training session lasted at least 15 minutes longer than originally scheduled due to high audience participation. Similarly, the pre- and post-tests also took longer than expected (45 minutes instead of 30), largely because of the participants’ high motivation levels. These analysts demonstrated a strong desire to accurately identify the posts, even though they knew their responses would remain anonymous. They also provided remarkably detailed and thoughtful responses to all qualitative sections.

4.8 Conclusions

4.8.1 Limitations

This work has several limitations. First, the sample size is relatively small. While this allowed us to gather detailed qualitative feedback and analysis, further research is needed to validate these results and generalize them to other populations. Second, the participants were government analysts who were generally more educated than the average American. All participants had at least some college education, with most holding a bachelor’s degree or higher. However, they were motivated by the fact that they would continue earning their full-time salary while participating in this training and would receive training credits. The types of analysts who enrolled in the training were highly motivated, resulting in very detailed and thoughtful qualitative responses. Finally, the two interactive, in-person training sessions took over an hour to administer, which is impractical for scaling to the general population. Nonetheless, this work demonstrates the potential usability of similar training sessions among motivated individuals already skilled in detecting misinformation.

4.8.2 Contributions

In a study conducted with motivated government analysts, we examined the effectiveness of media literacy training on misinformation detection and the participants’ willingness and ability to counter it. Previous research on media literacy has shown conflicting results regarding its effectiveness, and there has been little to no research on its usefulness in training individuals to combat misinformation. Overall, we found some, though limited, evidence that participants’ truth discernment improved in the post-test without lowering their confidence or increasing skepticism towards all news. We identified more substantial evidence that the training effectively increased participants’ willingness and likelihood to counter misinformation on social media.

Additionally, we conducted valuable qualitative work on why some individuals choose to counter misinformation while others do not. Understanding the reasons behind people’s willingness or reluctance to intervene, along with demonstrating that training can boost individuals' willingness to act, is crucial for determining how to improve social corrections and other user-driven countermeasures. For example, closeness emerged as a significant factor. People may feel more comfortable addressing misinformation with those they are close to, believing they can make a more substantial impact or at least should attempt to do so among loved ones. Verified accounts or those with large followings were considered more important to counter. The ease of reporting a post or user was also emphasized. By knowing these factors, social media companies can better design their platforms. For example, platforms could encourage more reporting by improving reporting functions and promote more social corrections by highlighting posts from closer contacts.

Chapter 5

Characterizing Platform and Government Countermeasures

In addition to research on user-based countermeasures, increasing attention has been given to misinformation mitigation efforts at both the platform [315] and government levels [248, 314]. The literature consistently shows that public opinion plays a critical role in shaping policy implementation and effectiveness [63]. A review of recent work in the misinformation space demonstrates that public support is a key factor to consider when developing successful interventions [89, 163, 164]. Understanding why people support or oppose specific countermeasures is therefore essential.

This chapter investigates several factors that have been previously identified as relevant to public policy support across various domains [118], including climate change initiatives [42, 136] and public health interventions [51, 87]. These features are fairness, intrusiveness, and effectiveness, and they are particularly relevant to misinformation interventions, as concerns over censorship and equity frequently arise among social media users [89, 295].

The primary research question for this chapter is: What types of countermeasures are supported by the general public and why? More specifically,

1. To what extent does a misinformation intervention's perceived **fairness**, **intrusiveness**, and **effectiveness** predict **support**?
2. How do the attributes people consider when forming preferences change due to the **implementer** (social media companies vs. governments) of the intervention?
3. What demographic factors, if any, are related to opinions on these topics?

5.1 Introduction

While the extent of misinformation's impact remains a subject of debate [11, 61, 94], its documented consequences are far-reaching. Research has linked misinformation to the erosion of democratic norms and institutions [94, 286], as well as the proliferation of violent and extreme conspiracy theories [84, 100, 252]. In response, a growing body of research has focused on developing effective and practical misinformation countermeasures, including social corrections

[28], warning labels [196], accuracy prompts [228], and various government regulations [238].

Several review articles have evaluated the extensive literature on misinformation interventions [79, 131, 314]. Interventions are typically evaluated by their effectiveness in reducing the creation, spread, or belief in misinformation by evaluating methods to improve truth discernment and corresponding behavior. However, to practically counter misinformation at scale, the public needs to trust and engage with the interventions.

Indeed, there have been calls for public participation in misinformation countermeasures [89, 164]. Research consistently shows the significant impact public opinion can have on the development and efficacy of public policy [63]. Furthermore, social media platforms are unlikely to implement unpopular countermeasures, as they are responsive to the desires of their users and any potential revenue implications [184]. Public acceptance of various countermeasures is a necessary yet understudied component of misinformation interventions. Therefore, we are motivated to address *why* people support or do not support certain inventions with the assumption that support is required for engagement.

For this chapter, we surveyed 1,010 residents of the United States who use social media at least once a week. This survey represents the second half of the one described in Chapter 3. Participants were asked to rate their support for various potential interventions on a Likert scale ranging from 1 to 5. Half of the participants were asked as if the government were implementing these policies, while the other half were asked as if the social media companies were doing so. Additionally, they were asked to rate how effective, fair, and intrusive they believed each intervention to be. We included ten interventions designed to span most of the categories outlined in Chapter 1.

5.2 Related Work

Recent research has examined the relationship between various personal attributes, including partisanship, trust in institutions, and previous exposure to misinformation or interventions, with support for interventions [189, 255]. However, the qualities of the countermeasures themselves that predict preferences remain unaddressed. Previous studies on the features influencing support for climate change policies identified three primary attributes: fairness, intrusiveness, and effectiveness [136]. These attributes are directly relevant to the misinformation context due to concerns about the potential infringement on free speech rights and the possible disproportionate impact of countermeasures on specific groups, such as Republican social media users [73, 238, 244].

Free speech is a fundamental right and value in the U.S., essential for a functioning democracy. However, Americans are divided on whether free speech or the restriction of false content should be prioritized [199]. We generally expect Americans to want countermeasures that protect free speech as much as possible while limiting the distribution and impact of false or misleading information. In other words, people aim to maximize effectiveness and fairness while minimizing the intrusiveness of these measures. Countermeasures involving user participation tend to limit the need for content removal and algorithmic manipulation, which arguably threaten free speech rights the most and provide the least transparency. Therefore, we believe that these attributes are timely and relevant for the implementation and communication strategies regarding misinformation mitigation.

Our first research question focuses on how perceived fairness, intrusiveness, and effectiveness of an intervention are related to the support of an intervention. We also explore whether user preferences differ based on whether the intervention is implemented by social media platforms or the government. With respect to countering misinformation, American citizens have expressed more concern about the government infringing on their free speech rights than private companies [199]. Therefore, perceptions of fairness, intrusiveness, and effectiveness may vary depending on which entity is responsible for these mitigation measures.

RQ1.1: To what extent does a misinformation intervention's perceived fairness, intrusiveness, and effectiveness predict support?

RQ1.2: How do the attributes people consider when forming preferences change due to the implementer of the intervention?

Next, we compare the general support, perceived fairness, perceived effectiveness, and perceived intrusiveness for each intervention.

RQ2: What is the average and variance in support, perceived fairness, perceived intrusiveness, and perceived effectiveness for each intervention?

Furthermore, certain segments of the U.S. population on social media may be more or less accepting of misinformation interventions due to different attributes. Understanding demographic and partisan differences in intervention support informs public messaging and intervention design, as well as larger trends in values involved in policy support judgments.

RQ3.1: How strongly do demographic differences predict support for misinformation interventions?

RQ3.2: Does support depend on different attributes for different demographic groups?

This work will have implications for the public communication strategies that governments and social media companies will use to foster support and engagement with misinformation mitigation efforts. Furthermore, to the best of our knowledge, this study is the first attempt to directly measure public support for a wide range of possible intervention strategies implemented by government entities and social media companies.

5.2.1 Intervention Selection

We examined ten interventions that could be implemented by either a social media platform or a government entity in this study. These interventions were selected to represent a broad range of possible countermeasures. Existing review articles in the misinformation intervention area have categorized interventions similarly; however, there is no common typology [79, 122, 131, 314]. After reviewing previous categorizations and drawing from my intervention categorization in

Chapter 1, we included the six general categories of countermeasures that could apply to both platforms or governments: content distribution, content moderation, account moderation, content labeling, media literacy, and institutional measures. User-based measures were excluded. We identified 1-2 representative interventions per category to present to study participants. Participants were told in advance whether the implementer of the intervention was social media platforms or government entities, explicitly mentioning that the determiner of what misinformation is would fall on the intervention implementer (e.g., platforms could fact-check internally or use an external, independent organization). The interventions are described in Table 5.1.

Table 5.1: Selected misinformation interventions. Text in [brackets] was included when the specified implementer was government.

	Category	**Intervention**	**Ref(s)**
1.	Content distribution/ friction	[Require social media companies to] temporarily delay users posting content the user did not open or spent less than a certain amount of time viewing.	[151, 231, 236]
2.	Content distribution/ advertising policy	[Require social media companies to] put all advertising through a fact-checking process.	[34, 131]
3.	Content moderation/ alg. downranking	[Require social media companies to] de-emphasize posts that are verified to contain misinformation.	[34, 109]
4.	Content moderation/ content removal	[Require social media companies to] remove posts verified to contain misinformation.	[34, 143]
5.	Account moderation/ account removal	[Require social media companies to] permanently ban users who post misinformation a certain number of times.	[30, 241]
6.	Content labeling/ misinfo. disclosure	[Require social media companies to] notify users if they posted content verified to contain misinformation.	[79, 160]
7.	Content labeling/ fact-check labels	[Require social media companies to] publicly label posts verified to contain misinformation with information about and from verified sources.	[222, 315]
8.	Media literacy	Invest in digital media literacy and promote educational content about detecting misinformation on and offline.	[121, 162, 251]
9.	Institutional measures/ media support	Promote and invest in local media, which is thought to be most in tune with local norms, culture, and context.	[32, 52]
10.	Institutional measures/ data sharing	[Require social media companies to] regularly release data and/or internal research reports about misinformation prevalence, spread, and mitigation to the public and outside researchers.	[22, 34, 52]

5.3 Data and Methods

The survey questions associated with this Chapter were included in the same survey referenced in Chapter 3. The same ethics information and sampling plan apply. See Section 3.2 for more details.

5.3.1 Platform and Government Survey Questions

After the informed consent, qualifying questions, and individual behavior and opinion questions, participants progressed to the second half of the survey. Each participant was randomly assigned to see interventions implemented by either the government or social media companies. Each participant saw a random subset of 8 (of 10) interventions. They received the following instructions:

> This next section concerns policies that could be implemented by [the government / social media companies] to limit the spread and influence of misinformation.
>
> This means that, when applicable, **the [government / social media companies] would determine what misinformation is** when implementing the intervention(s).

Next, participants were instructed to evaluate one potential policy at a time, selected from those described in Table 5.1. They were first asked how much they support or oppose the proposed policy on a five-point Likert scale, which ranged from “Strongly support” to “Strongly oppose.” After that, participants were asked about perceived effectiveness, fairness, and intrusiveness. Each participant was asked about these three factors in a random order to mitigate potential question-order biases. This random order was maintained for all policies presented to each participant. The three factor questions are displayed below:

> In your opinion, would the proposed policy be **effective** or **ineffective** in reducing misinformation on social media? *[Very effective, Somewhat effective, Neither effective nor ineffective, Somewhat ineffective, Very ineffective]*
>
> In your opinion, would the proposed policy be **fair** or **unfair** when applied to different types of social media users? *[Very fair, Somewhat fair, Neither fair nor unfair, Somewhat unfair, Very unfair]*
>
> In your opinion, would the proposed policy be **intrusive** or **unintrusive** on the experience of social media users? *[Very intrusive, Somewhat intrusive, Neither intrusive nor unintrusive, Somewhat unintrusive, Very unintrusive]*

The survey design and the phrasing of the factor questions were modeled off of prior work [136]. In particular, the phrasing of the fairness question was designed in order to frame it in terms of perceived distributional fairness (effects on various groups) rather than personal fairness (effects on oneself) [42, 118, 147]. This choice was deliberate, as not every participant may inherently consider the same definition of fairness when evaluating the fairness of different policies. Furthermore, a meta-analysis found that perceived distributional fairness had a significantly stronger effect on support levels for climate change policies compared to perceptions of personal fairness [42].

5.3.2 Measures

Support for intervention(s) We asked participants to rate a subset of interventions 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 effectiveness of intervention(s) We asked participants to rate a subset of interventions 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).

Perceived fairness of intervention(s) We asked participants to rate a subset of interventions 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 asked participants to rate a subset of interventions 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).

5.3.3 Pre-Registered Analysis Plan

In our pre-registration (`https://osf.io/b2yjt/`), we included the primary research questions and analysis plan for this study.

RQ1: Predicting Support

To model support levels as a function of factor perceptions and implementer, we initially planned to run a multilevel model to account for random effects (i.e., random slope and intercept) of interventions and participants, as each participant saw 8 of the 10 selected interventions, drawn randomly, and multiple participants rated each intervention. We also pre-registered that if this model did not converge, we would fit an OLS regression model with robust standard errors clustered on participants and interventions. Since the multilevel model did not converge, the model output reported in this article is an OLS regression with robust standard errors clustered on participants and interventions. We additionally ran planned robustness checks by including participants who responded to a part of the survey but did not complete it. This does not change the direction or significance of the effects found in the primary model.

We calculated adjusted fractional Bayes factors with Gaussian approximations for the primary models using the BFPack R package [205]. We report BF10 for each estimate where the alternative hypothesis is directional based on the sign of the estimate (i.e., $b < 0$, $b > 0$) and the null hypothesis is $b = 0$. Thus, if $BF > 1$, the evidence is more consistent with the alternative hypothesis; if $BF < 1$, the evidence is more consistent with the null hypothesis.

RQ2: Descriptive Analyses

We described how we would also report descriptive analyses of the average and spread of support for each intervention, as well as for the perceptions of effectiveness, fairness, and intrusiveness.

RQ3: Individual Differences

We pre-registered our planned regression modeling analysis to determine how ratings for support, effectiveness, fairness, and intrusiveness vary based on individual differences of the participant. We included standard demographic variables (gender, age, income, education), partisanship (party id and political ideology), and self-reported exposure to misinformation. The question of misinformation exposure was asked in part 1 of the survey on individual behaviors and opinions (see Chapter 3).

5.3.4 Adhoc Analysis

To enhance our analysis of individual differences in support and perceptions of interventions, we conducted one-way ANOVA tests comparing average support, perceived fairness, perceived effectiveness, and perceived intrusiveness across categories for each demographic variable measured categorically (i.e., partisanship, gender, age, income, education, ideology, and misinformation exposure frequency).

After identifying the significance of the partisanship and gender demographic factors, we incorporated interactions between these factors, implementer, and perceptions of fairness, effectiveness, and intrusiveness to predict support in an ad hoc analysis. We also ran the same model with gender and political ideology instead of partisanship.

5.4 Results

5.4.1 Factors that Influence Support for Interventions

RQ1: To what extent does a misinformation intervention's perceived fairness, intrusiveness, and effectiveness predict support? How do the attributes people consider when forming preferences change due to the implementer of the intervention?

Figure 5.1 and Table 5.2 show the regression results for RQs 1.1 and 1.2. Perceived fairness is most strongly associated with support ($\beta = 0.624, SE = 0.016, p < 0.001$), followed by perceived effectiveness ($\beta = 0.302, SE = 0.015, p < 0.001$) and intrusiveness ($\beta = -0.065, SE = 0.010, p < 0.001$). These effects are moderated by the implementer, though implementer on its own is not significant in the model (β = 0.123, SE = 0.089, p = 0.167, reference level: social media company). Fairness is less associated with support when the implementer is government than social media platforms ($\beta = -0.080, SE = 0.024, p < 0.001$), while intrusiveness ($\beta = -0.036, SE = 0.015, p = 0.015$) and effectiveness ($\beta = 0.077, SE = 0.022, p < 0.001$) are more strongly associated.

5.4.2 Overall Support and Perceptions of Interventions

RQ2: What is the average and variance in support, perceived fairness, perceived intrusiveness, and perceived effectiveness for each intervention?

Table 5.2: OLS regression predicting support for a misinformation intervention as a function of perceived fairness, perceived effectiveness, perceived intrusiveness, and implementer (reference level: social media company) with robust standard errors clustered on participant and intervention. Model 1 is the main model. Model 2 is the robustness check that includes responses from all participants who responded to at least part of the survey.

	Model 1	Model 2
	Estimate (Standard error)	
(Intercept)	0.586^{***}	0.582^{***}
	(0.060)	(0.060)
	BF > 100	BF > 100
Implementer	0.123	0.123
	(0.089)	(0.089)
	BF = 0.053	BF = 0.054
Perceived fairness	0.624^{***}	0.626^{***}
	(0.016)	(0.016)
	BF > 100	BF > 100
Perceived effectiveness	0.302^{***}	0.300^{***}
	(0.015)	(0.015)
	BF > 100	BF > 100
Perceived intrusiveness	-0.065^{***}	-0.064^{***}
	(0.010)	(0.010)
	BF > 100	BF > 100
Implementer × perceived fairness	-0.080^{***}	-0.083^{***}
	(0.024)	(0.024)
	BF = 5.91	BF = 9.34
Implementer × perceived effectiveness	0.077^{***}	0.080^{***}
	(0.022)	(0.022)
	BF = 10.31	BF = 17.93
Implementer × perceived intrusiveness	-0.036^{*}	-0.036^{*}
	(0.015)	(0.015)
	BF = 0.42	BF = 0.43
Observations	8,071	8,102
R^2	0.760	0.761
Adjusted R^2	0.760	0.760

Note: $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$

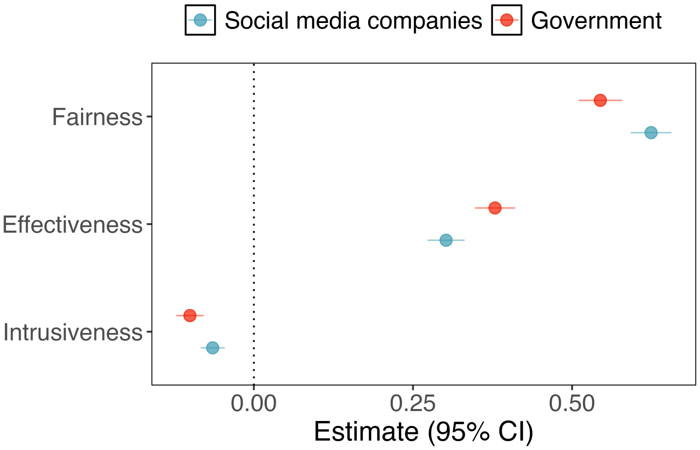


Figure 5.1: Estimate and 95% CI of the effect of perceived fairness, effectiveness and intrusiveness on support depending on implementer.

Figure 5.2 contains the estimate and 95% CI for support, perceived fairness, perceived effectiveness, and perceived intrusiveness for each intervention by each implementer (government and social media company). Participants were more supportive of interventions in the content labeling category and less supportive of those in the content distribution or moderation categories.

5.4.3 Individual Differences in Support and Perceptions of Interventions

Next, we investigate individual differences in support and perceived attributes of misinformation interventions (full regression output in Table 5.3). We do not find significant differences in support for interventions across age groups, education level, or frequency of previous exposure to misinformation. However, we do find substantial differences in both support and perception of interventions among partisan and gender groups.

Unsurprisingly, we find Democrats support interventions more than Independents/other ($\beta = -0.284$, $SE = 0.074$, $p < 0.001$) and Republicans ($\beta = -0.285$, $SE = 0.107$, $p = 0.008$). We also find Independents/other and Republicans perceive interventions as less fair and effective than Democrats ($p < 0.01$ for all), while only Independents/other perceive interventions as more intrusive than Democrats ($\beta = 0.212$, $SE = 0.066$, $p < 0.01$). These results are robust to separating the "Independent" and "Other/unaffiliated" categories and mapping partisan categories to corresponding numeric values (see Appendix G). Similarly, liberal leaning participants tend to support interventions more than conservative leaning participants ($\beta = -0.295$, $SE = 0.037$, $p < 0.001$). Liberal ideology is also associated with perceiving interventions as more fair, more effective, and less intrusive ($p < 0.001$ for all).

We find men support interventions less than women on average ($\beta = -0.199$, $SE = 0.054$, $p < 0.001$). Men perceive countermeasures as less fair, less effective, and more intrusive than women on average as well ($p < 0.05$ for all). Moreover, people with higher incomes tend to support interventions more ($\beta = 0.031$, $SE = 0.016$, $p = 0.047$) and perceive them as more fair ($\beta = 0.045$, $SE = 0.016$, $p < 0.01$). Finally, older participants perceive interventions as

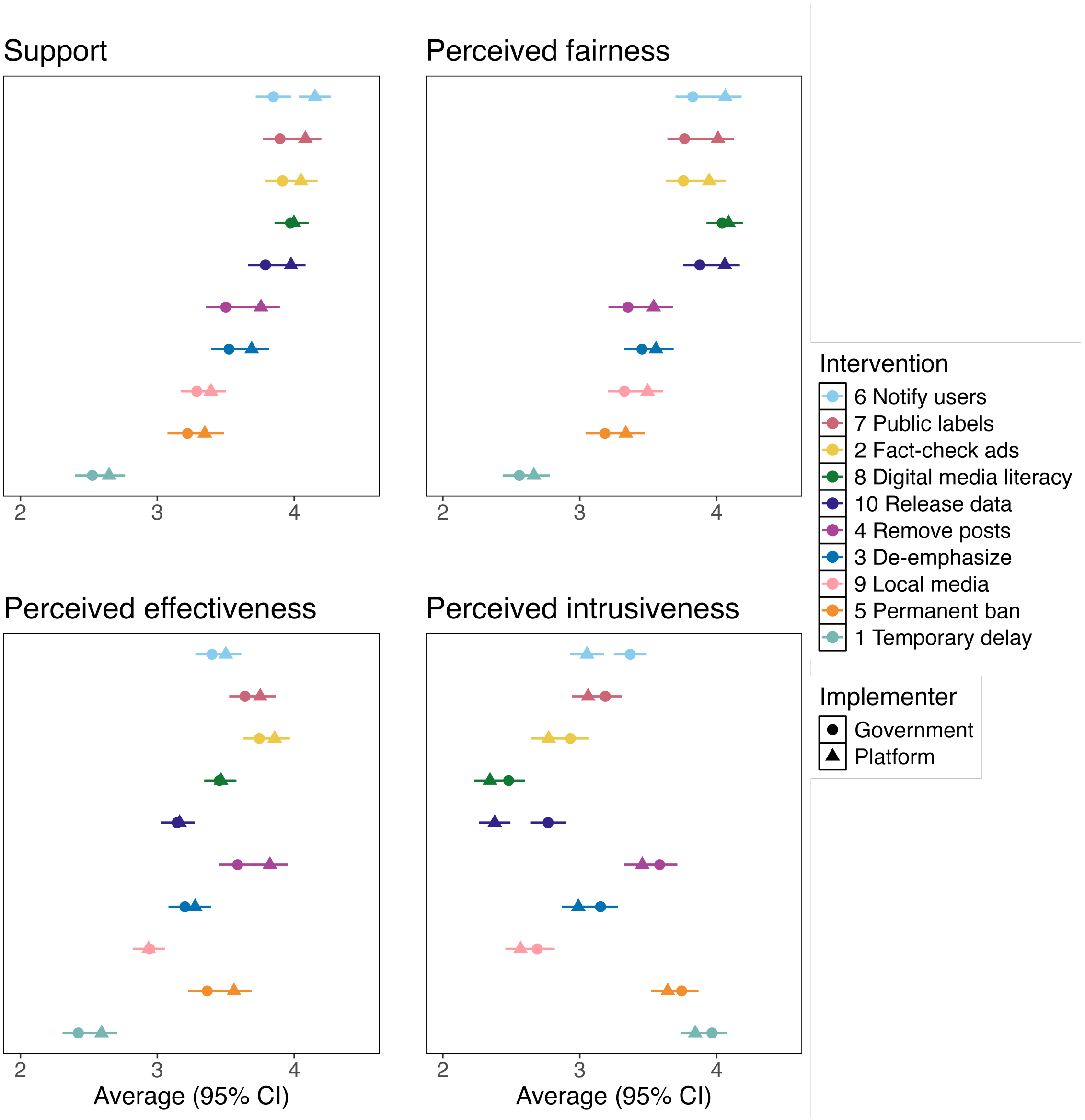


Figure 5.2: Average and 95% CI support, perceived fairness, perceived effectiveness, and perceived intrusiveness for each intervention (1-10) and implementer (government and platform) on a 1-5 Likert scale.

more intrusive ($\beta = 0.038$, $SE = 0.018$, $p < 0.05$) and more fair ($\beta = 0.044$, $SE = 0.021$, $p < 0.05$) than younger participants. Education level and previous exposure to misinformation are not associated with support level or any perceptions of interventions.

To better quantify these differences, when we compare the average Likert scores for support, we find that Republicans and Independents/others support misinformation interventions 23.8% and 15.3% less than Democrats, respectively. They also perceive these interventions as less fair (22% and 14.1%) and less effective (17.8% and 12.1%). For gender, we find that men support misinformation countermeasures 6.7% less than women on average. Furthermore, men rate interventions as 5.1% less fair, 6.9% less effective, and 3.8% more intrusive than women.

To complement our pre-registered regression analysis, we ran one-way ANOVA tests comparing average support, perceived fairness, perceived effectiveness and perceived intrusiveness across categories for each demographic variable (see Appendix H for the full results). We found statistically significant differences in support for gender, partisanship, ideology, and education ($p < 0.05$ for all). Unlike in the regression analysis, income groups do not differ in support.

Figure 5.3 shows the average support level, perceived fairness, perceived effectiveness, and perceived intrusiveness broken down by partisanship and gender. Notably, gender, partisanship and ideology are the only variables that differ across all outcomes in both regression and ANOVA analyses (except the ANOVA for gender and intrusiveness). We exclude ideology from Figure 5.3 for conciseness as partisanship and political ideology are strongly associated (see Table 5.4). Analogous figures for the remaining demographic variables can be found in Figure 5.4.

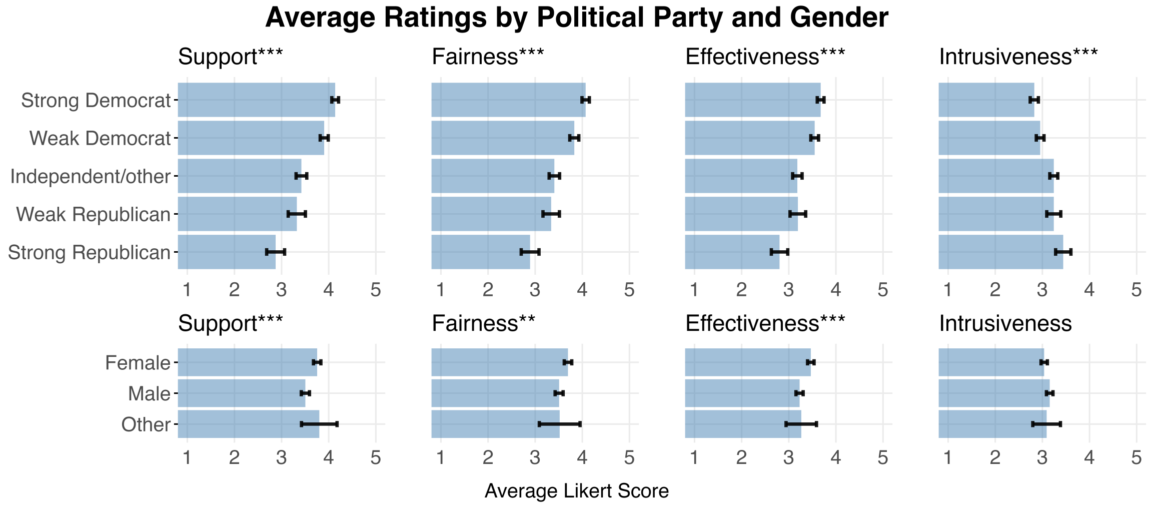


Figure 5.3: Average ratings by political party and gender. One-way ANOVA tests were run on each grouping, with stars indicating the level of significance: $p < 0.05$*, $p < 0.01$**, and $p < 0.001$***

Figure 5.5 recreates Figure 5.3 but with average Likert values and 95% confidence intervals calculated separately based on the implementer. It becomes clearer that the divergence in average values among strong Republicans stems from a significant difference in perceptions and support levels between government and platform interventions, with government interventions scoring more poorly. Analogous figures for each intervention can be found in Appendix H.

Table 5.3: OLS regressions predicting average support, perceived fairness, perceived effectiveness, and perceived intrusiveness of interventions as a function of age (18-24 to 65+ age brackets mapped to numeric values 0 to 5), gender (reference level: Female), partisanship (reference level: Democrat, combined "Other" party with "Independent"), political ideology (very liberal to very conservative categories mapped to numeric values 0 to 4), highest education level (less than high school diploma to Doctorate or Professional Degree categories mapped to numeric values 0 to 6), income (less than $20,000 to over $200,000 income brackets mapped to numeric values 0 to 7), frequency seeing misinformation (never to very often mapped to numeric values 0 to 4).

	Dependent variable:			
	Avg. support rating	Avg. fairness rating	Avg. effectiveness rating	Avg. intrusiveness rating
Age	0.020	0.044*	−0.023	0.038*
	(0.021)	(0.021)	(0.019)	(0.018)
Male : Female	−0.199***	−0.137*	−0.214***	0.102*
	(0.054)	(0.055)	(0.050)	(0.048)
Other (e.g., non-binary) : Female	−0.114	−0.296	−0.325*	0.139
	(0.173)	(0.176)	(0.161)	(0.154)
Education	0.013	0.003	0.006	0.016
	(0.022)	(0.022)	(0.020)	(0.019)
Income	0.031*	0.045**	0.012	−0.005
	(0.016)	(0.016)	(0.015)	(0.014)
Independent / Other : Democrat	−0.284***	−0.270***	−0.218**	0.212**
	(0.074)	(0.075)	(0.068)	(0.066)
Republican : Democrat	−0.285**	−0.294**	−0.208*	0.161
	(0.107)	(0.109)	(0.099)	(0.095)
Political Ideology	−0.295***	−0.263***	−0.184***	0.130***
	(0.037)	(0.037)	(0.034)	(0.033)
Misinformation Exposure	−0.014	−0.007	−0.047	0.038
	(0.027)	(0.027)	(0.025)	(0.024)
Constant	4.236***	4.038***	3.997***	2.496***
	(0.122)	(0.124)	(0.114)	(0.109)
Observations	1,010	1,010	1,010	1,010
R^2	0.244	0.206	0.151	0.095
Adjusted R^2	0.237	0.199	0.144	0.087
Residual Std. Error (df = 1000)	0.834	0.847	0.775	0.744
F Statistic (df = 9; 1000)	35.809***	28.792***	19.798***	11.711***
Note:				*p<0.05; **p<0.01; ***p<0.001

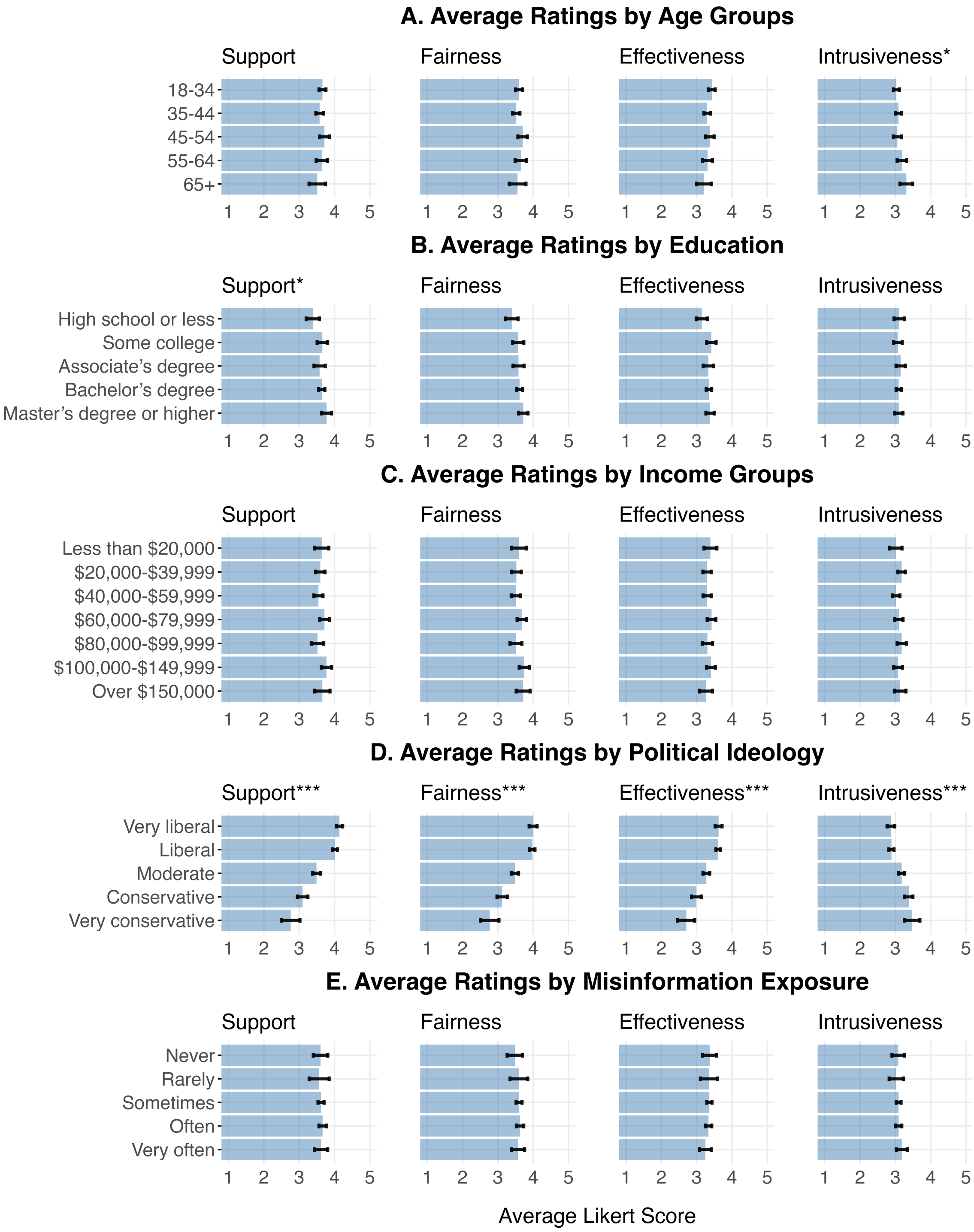


Figure 5.4: Average ratings by age, education, income, political ideology, and misinformation exposure. One-way ANOVA tests were run on each grouping, with stars indicating the level of significance: $p < 0.05$*, $p < 0.01$**, and $p < 0.001$***

Table 5.4: Contingency table of partisan and political ideology groups. Fisher's test comparing party and ideology categories ($\chi^2 = 1072, p < 0.001$).

	Republican	**Independent/Other**	**Democrat**
Very liberal	4	17	148
Liberal	3	37	265
Moderate	20	206	46
Conservative	138	40	9
Very conservative	71	3	3

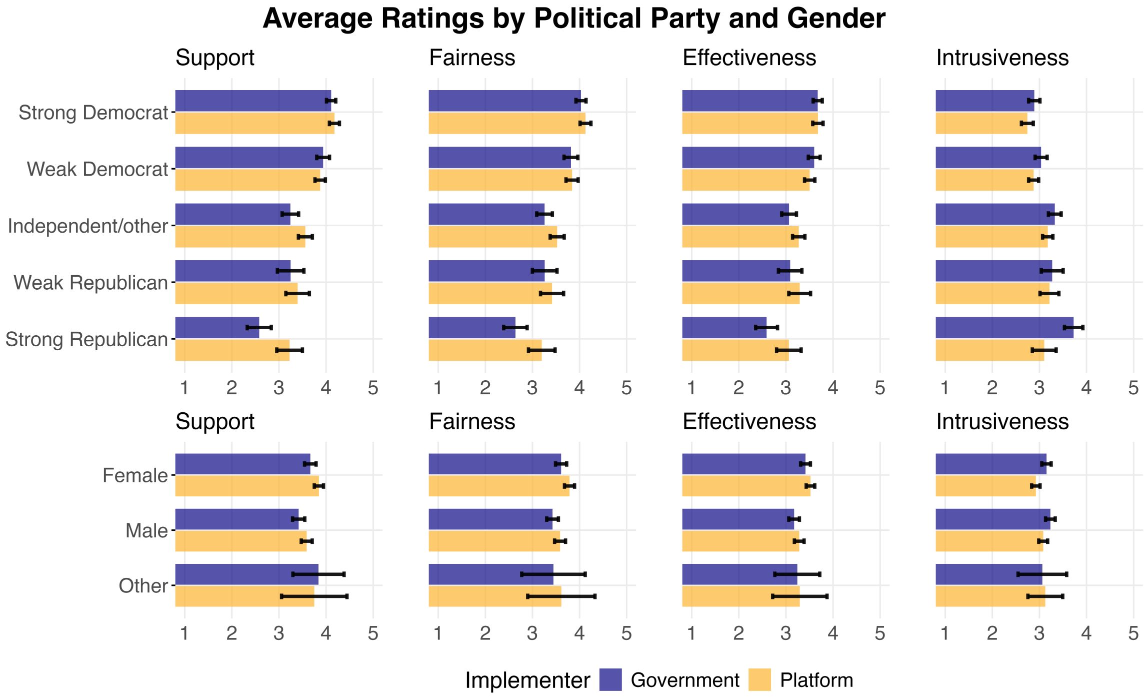


Figure 5.5: Average Likert ratings and 95% confidence intervals by political party and gender, separated by intervention implementer.

5.4.4 Partisanship Influences Features that Predict Support

From our analysis of individual differences in support, we identified partisanship and gender to examine further. We performed an ad hoc analysis to examine how gender and partisanship interact with the implementer of the intervention and perceived attributes to predict support (i.e., add demographic variables to the model used in RQ1); see Table 5.5. We find that partisanship interacts with implementer and fairness, with Republicans caring more about fairness than Democrats ($\beta = 0.061, SE = 0.030, p = 0.040$). There is also a larger difference in support for interventions implemented by government versus social media companies for Republicans and Independents than for Democrats ($\beta = -0.094, SE = 0.039, p = 0.017$; $\beta = -0.108, SE = 0.036, p = 0.003$), where interventions implemented by governments are less supported.

After accounting for partisanship and gender, the implementer becomes a significant factor in the model. Specifically, platform-led interventions receive greater support than government-led ($\beta = 0.213, SE = 0.094, p = 0.024$). As shown in Figure 5.2, interventions implemented by companies are generally supported more and are overall perceived more positively across the three factors. In addition, we ran the same model with political ideology included instead of partisanship (see Appendix G). We again find a significant interaction between political ideology and perceived fairness, where more conservative-leaning participants weigh fairness more than liberal-leaning participants ($\beta = 0.028, SE = 0.010, p = 0.005$).

5.5 Discussion

In this work, we surveyed American social media users to examine public acceptance of interventions against misinformation implemented by the government and social media companies. We found that belief in fairness was most strongly associated with support for an intervention, followed by effectiveness, and finally intrusiveness. Fairness was more of a concern when the implementer was social media companies than the government, while effectiveness and intrusiveness were more salient when the government was the implementer (Figure 5.1). However, in general, the same intervention implemented by social media companies received more support, was perceived as fairer and more effective, and was viewed as less intrusive than when implemented by the government (Figure 5.2). These findings may reflect public attitudes towards businesses and government, where companies are more trusted to address misinformation in a timely manner at scale. Alternatively, people may believe that social media companies have a greater responsibility to address misinformation than the government.

Our results further indicate that people desire agency and transparency in misinformation interventions, echoing findings from Saltz et al. [255] and research from other policy contexts [85, 118]. They support interventions that provide information to users that they can use when deciding how to interact with certain content, such as notifying them if they have posted misinformation, adding public labels to content containing misinformation, implementing digital media literacy programs, and requiring platforms to release of data or internal research reports related to misinformation. There was also strong support for holding advertising accountable through fact-checking. People were generally less supportive of interventions that involve re-

Table 5.5: Adhoc analysis: OLS regression predicting support for a misinformation intervention as a function of perceived fairness, perceived effectiveness, perceived intrusiveness, implementer (reference level: social media company), gender (reference level: Female), and partisanship (reference level: Democrat, with Independents combined with "Other") with robust standard errors clustered on participant and intervention.

	Dependent variable: Support	
	Estimate	Std. Err.
Implementer (platform : government)	0.213*	(0.094)
Perceived fairness	0.564***	(0.025)
Perceived effectiveness	0.303***	(0.022)
Perceived intrusiveness	−0.063***	(0.015)
Gender (male : female)	−0.231**	(0.089)
Gender (other : female)	0.706*	(0.296)
Party (Independent/Other : Democrat)	−0.251*	(0.111)
Party (Republican : Democrat)	−0.242*	(0.113)
Implementer x Perceived fairness	−0.086***	(0.024)
Implementer x Perceived effectiveness	0.066**	(0.022)
Implementer x Perceived intrusiveness	−0.034*	(0.014)
Implementer x Gender (male : female)	0.010	(0.030)
Implementer x Gender (other : female)	0.188	(0.107)
Implementer x Party (Independent/Other : Democrat)	**−0.108****	(0.036)
Implementer x Party (Republican : Democrat)	**−0.094***	(0.039)
Perceived fairness x Gender (male : female)	0.047	(0.024)
Perceived fairness x Gender (other : female)	−0.087	(0.060)
Perceived fairness x Party (Independent/Other : Democrat)	0.038	(0.029)
Perceived fairness x Party (Republican : Democrat)	**0.061***	(0.030)
Perceived effectiveness x Gender (male : female)	−0.014	(0.022)
Perceived effectiveness x Gender (other : female)	−0.037	(0.068)
Perceived effectiveness x Party (Independent/Other : Democrat)	0.018	(0.026)
Perceived effectiveness x Party (Republican : Democrat)	−0.021	(0.027)
Perceived intrusiveness x Gender (male : female)	0.020	(0.015)
Perceived intrusiveness x Gender (other : female)	−0.059	(0.048)
Perceived intrusiveness x Party (Independent/Other : Democrat)	−0.009	(0.018)
Perceived intrusiveness x Party (Republican : Democrat)	−0.021	(0.019)
Constant	0.889***	(0.096)
Observations	8071	
R^2	0.766	
Adjusted R^2	0.766	
Note:	*p<0.05; **p<0.01; ***p<0.001	

moving or de-emphasizing posts identified as containing misinformation. While banning users who repeatedly post misinformation is also largely not supported, many believed that it would be relatively effective. Belief in effectiveness is simply not enough to support certain interventions that are considered unfair or intrusive. These results are consistent with the literature in other policy areas, which finds that the public generally prefers informational interventions over more restrictive measures even though they are often less effective [87, 123].

Interestingly, two of the least supported interventions do not directly threaten free speech or involve any censoring activity. Promoting and investing in local media was perceived to be largely ineffective. It may be that this intervention was too vague for participants to envision how it could help mitigate misinformation. Finally, the least popular intervention by a large margin was temporarily delaying users when attempting to post content they did not or barely viewed. It was rated as the least fair and effective and the most intrusive. This result is fairly unexpected considering that some platforms currently implement this intervention, including Facebook and X [74, 236]. Social media affords instantaneous communication and content, which could drive impatience for even the slightest inconvenience or delay (e.g., commercial breaks [60]). Therefore, when employing nudge-based approaches like accuracy prompts [228] or temporary delays in posting, it is imperative to minimize intrusiveness in the user experience, such as through design choices.

The analysis of individual differences in support for and perceptions of interventions revealed that men tend to support misinformation interventions less than women, and that Republicans and Independents support them less than Democrats. In addition to supporting interventions less, men and Republicans viewed them as less fair, less effective and more intrusive than women and Democrats, respectively. The gender gap reflects broader trends of women supporting more (government) regulations than men across policy domains [2]. The gap in support is, in part, explained by differences in perceptions of the proposed interventions. Future work should examine how perceived features of policies interact with other factors like emotional reactions and issue awareness to predict differences in support between genders [258].

Furthermore, the partisan differences align with findings from Saltz et al. [255]. Previous studies have shown that Republicans are more likely to perceive interventions as biased against them [255, 295]. A 2022 Pew Research poll found that approximately 70% of Republicans believe that major technology companies favor the views of liberals over conservatives, while only 22% of Democrats say they believe companies favor conservatives over liberals [295]. It is likely that because Republicans believe that they are more likely to be censored for their viewpoints, they perceive interventions as less fair and are, therefore, less supportive of all interventions potentially employed by companies or the government. Whether this perception is accurate or not, social media companies must work on rebuilding trust among all their users.

Fairness emerged as the most significant predictor of support for interventions, and it varied across demographic groups, with fairness being especially important for Republicans. This emphasis on fairness is not unique to misinformation and social media policies. Across a variety of policy contexts, including health, environment, and transportation, perceived fairness and effectiveness are often among the most predictive factors associated with support [31, 42, 51, 118]. In fact, a systematic review found that enhancing the communication of a potential policy's effectiveness to participants can boost support levels by 4%, a small but meaningful increase [243]. Fairness also plays a significant role and has been found to be the most important factor in a variety of public health interventions, such as those promoting healthy food [51], as well as support for climate change policies [42].

Intrusiveness is also a related factor, but it is not surprising that it holds less importance in the social media space than in other policy contexts. Previous studies have shown that when interventions are perceived to more directly impact an individual's personal choices or daily life, like increased costs associated with owning a car [147], they tend to be unpopular [87, 123, 136]. It may be that the potential intrusiveness of social media policies on the user experience does not result in as severe of consequences for the typical individual as those in other policy areas, which could impact access to food, health care, transportation, and more.

5.6 Conclusions

5.6.1 Limitations

While this study is among the first to analyze the factors associated with support for misinformation interventions, several limitations could be addressed in future work. First, we focused our survey on active social media users, as those individuals would be the most affected by any potential policies and the most familiar with current interventions. We also only included ten interventions to limit the length of the survey, allowing us to focus more on the factors behind support. However, several new and emerging interventions were not included. For example, X's (formerly Twitter) Community Notes program has been relatively successful at increasing the volume of fact-checks and boosting trust in misinformation flags by using crowd-sourcing misinformation detection and labeling [71, 90]. However, reports of its overall effectiveness in reducing engagement with misleading content are mixed [71, 148]. Future research should supplement these results by surveying a wider range of interventions.

Additionally, we focused exclusively on effectiveness, fairness, and intrusiveness. More factors (e.g., transparency) should be considered in future surveys about public acceptance of policies. Problem awareness may also be an important aspect to consider in future research [87, 118]. Finally, in the survey, we indicate that the implementer of the intervention is responsible for misinformation detection, which is an oversimplification. Future work should assess how people think misinformation should be detected.

5.6.2 Contributions

Collectively, our findings suggest that fairness is valued above intrusiveness and effectiveness when determining support for misinformation interventions, and it is especially critical for specific demographic groups like Republicans. When designing and implementing misinformation interventions, mitigating any possible disproportionate impacts on certain groups or individuals is critical. In addition, public messaging should emphasize why each intervention is needed and how they are being implemented fairly, in addition to providing recourse for users when necessary. Furthermore, there is more support for and positive perceptions of interventions deployed by social media companies than the government, which may reflect broader trends in institutional trust. Most likely, effective misinformation interventions require collaboration across institutions. However, broader support for company-implemented interventions can be leveraged in public communications and education.

Our analysis of support levels and perceived features of interventions highlights the importance of promoting user agency to garner widespread support. For example, platforms can allow users to engage with misinformation warnings and nudges behind interstitials rather than strictly and opaquely removing violating content. At the same time, interventions should be carefully designed and implemented to minimize disruption to the user experience. Overall, this work has important implications for designing misinformation interventions and messaging that will be positively received by social media users.

Chapter 6

Recommendations for Effective and Practical Countermeasures

While earlier chapters have characterized a wide range of countermeasures and sought to evaluate and improve those interventions, this chapter consolidates that prior work to establish a framework for developing effective and practical countermeasures. The objective of this chapter is to provide a comprehensive set of recommendations across the intervention landscape that researchers, companies, and policymakers can use.

To develop this framework, I integrate the research from previous chapters and evaluate interventions based on several critical features: **effectiveness, acceptance, effort level, cost,** and **political feasibility**. Additionally, an expert survey was conducted to gather professional opinions on these factors and incorporate them into this analysis.

The main research question for this chapter is: How can we identify which interventions are both practical and effective, and under what circumstances? More specifically:

1. What are the characteristics of different interventions, and how do users perceive them?
2. What features do effective and practical countermeasures have in common?
3. Can we develop a framework for evaluating interventions?

6.1 Introduction

Interventions are unlikely to be implemented unless they are both effective and have support from the affected users or the public, at least to some extent [117]. However, while effectiveness and acceptance may be seen as necessary conditions for implementation, they are not sufficient on their own. The literature on comparative policy analysis indicates that other critical factors are associated with the likelihood of a policy's implementation, whether at the platform or government level: the level of implementation effort, intervention cost, and overall political feasibility [168, 248].

In this chapter, the previously defined categories and sub-categories of countermeasures, as described in Section 1.4 and Appendix A, are operationalized into actual, implementable interventions by platforms and institutions. The operationalized interventions are outlined in Section

6.2.1 and compared and ranked based on five evaluative metrics: effectiveness, acceptance, effort level, cost, and political feasibility. The objective characteristics and the user perceptions of the operationalized interventions are discussed as they affect these evaluative metrics. This analysis uses a combination of sources, including existing literature, results from previous chapters, and findings from an expert survey conducted for this chapter.

The objective characteristics of the operationalized interventions are informed by the countermeasures categorization from Chapter 1 and the scoping literature review from Chapter 2. The user perceptions of the interventions, particularly user acceptance, are assessed by synthesizing the results from Chapters 3-5. However, since the literature has limited values for effort level, cost, and feasibility, the five main evaluative metrics are primarily assessed based on the expert survey results.

6.2 Related Work

The misinformation countermeasures discussed in this dissertation have been categorized into general types, such as content distribution and account moderation (see Section 1.4.2). These categories were developed due to their shared key characteristics and features. Appendix A defines each of these eight general categories in more detail and outlines specific interventions for each.

6.2.1 Operationalized Interventions

This section examines how governments, platforms, and institutions can implement these measures in practice. The specific interventions were operationalized by reviewing the literature and platform policies to identify examples for each intervention type that have been implemented or proposed previously [6, 248].

Table 6.1 outlines operationalized interventions primarily implemented by platforms and includes implementations of the specific interventions from the Content Distribution, Content Moderation, Account Moderation, and Content Labeling categories. Some specific intervention sub-categories are used multiple times (e.g., advertising policy, algorithmic content moderation) or in combination with others when operationalized (e.g., fake news games and inoculation). Others are combined across categories (e.g., fact-checking and debunking are operationalized as fact-checking labels in the Content Labeling category).

Table 6.2 outlines the operationalized interventions primarily led by users and institutions and includes those focusing on the User-based Measures, Media Literacy and Education, Institutional Measures, and Generative AI/Other categories. All specific interventions defined in Appendix A are operationalized.

6.2.2 Objective Characteristics

Objective characteristics refer to the relatively static features that define how an intervention is implemented and by whom. These features include:

Table 6.1: Operationalized platform interventions.

Specific Intervention(s)	Operationalized Intervention
	1. CONTENT DISTRIBUTION
Redirection	**Redirection** - Redirect users to other content, such as official content or no content, when searching for something potentially problematic or harmful
Accuracy Prompts	**Accuracy Prompts** - Nudge or remind people to consider accuracy before posting or sharing content, such as by being asked to rate the accuracy of the headline
Friction	**Friction** - Temporarily delay users from posting content they did not open via a pop-up
Platform Alterations	**Platform Alterations** - Reduce the size or visibility of a post containing misinformation
Advertising Policy	**Ban Political Ads** - on social media platforms
Advertising Policy	**Fact-Check Ads** - on social media platforms
Content Distribution	**Limit Forwarding** - Cap the number of users one can forward a given message
Content Distribution	**Limit Resharing** - Remove share buttons on posts after several levels of sharing
	2. CONTENT MODERATION
Misinfo. Detection	**Content Removal** - Remove posts verified to contain misinformation.
Alg. Content Moderation	**Downranking** - De-emphasize posts in news feeds that contain misinformation
Alg. Content Moderation	**Algorithmic Changes** - uprank high-quality news sources and downrank low-quality ones
Alg. Content Moderation	**Virality Circuit Breakers** - Automatically flag certain fast-spreading and unverified content, triggering a brief halt on algorithmic amplification until the information is verified
Content Moderation	**User Control** - Give users more control over the algorithms powering their news feeds
	3. ACCOUNT MODERATION
Account Removal	**Account Suspensions** - Permanently or temporarily ban users who post misinformation or violate other platform policies a certain number of times
Account Removal	**Deplatforming** - Coordinated efforts among several social media companies to remove especially problematic or dangerous user accounts.
Shadow Banning	**Shadow banning** - Limit the spread of posts from certain policy-violating accounts without explicitly banning or suspending them.
Demonetization	**Demonetization** - The removal or restriction of monetization features for a user account found repeatedly violating a platform's policies.
	4. CONTENT LABELING
Fact-Checking, Debunking, Content Labeling	**Fact-Check labels** - Link information from 3rd party fact-checkers on misinformation posts
Crowdsourcing, Context Labels	**Crowdsourcing** - Labels generated by the general public rather than by professional fact-checkers, like Community Notes programs
Warning Labels	**Click-through Warning Labels** - Place misinformation posts behind click-through labels containing facts and context
Source Credibility Labels	**Source Labeling** - Labels indicating the reliability of news sources using, for example, NewsGuard labels
Source Credibility Labels	**Government Labels** - Label the accounts of government officials or state-run media

Table 6.2: Operationalized user-based and institutional interventions.

Specific Intervention(s)	Operationalized Intervention
	5. USER-BASED MEASURES
Reporting, Blocking	**Reporting** - Improved reporting functionality and transparency, to encourage users to report misinformation and harassment
Social Norms, Social Corrections, Retractions	**Social Norms** - The usage of social or community norms to encourage social corrections and self-corrections via platform changes
Social Norms	**Alter Platform Metrics** - Reward accuracy rather than engagement to discourage misinformation sharing
	6. MEDIA LITERACY AND EDUCATION
Media Literacy	**Digital Media Literacy** - Invest in and promote educational content on how to critically evaluate online information
Fake News Games, Inoculation	**Inoculation** - Inoculating people against misinformation with games or videos (For example, the Bad News Game)
	7. INSTITUTIONAL MEASURES
Media Support	**Media Support** - Promote and invest in local media, which is thought to be most in tune with local norms, culture, and context.
Media Support	**Journalism Support** - Supporting and training the next generation of journalists to engage in high-quality and independent reporting
Data Sharing	**Data Sharing** - Have social media companies regularly release data and/or internal research reports to the public and outside researchers.
Government Regulation	**Government Regulation** - Hold companies accountable for the content shared on their platforms. This could involve modifying Section 230, or regulating social media companies
Government Regulation	**Privacy Legislation** - The development of comprehensive privacy legislation, similar to Europe's GDPR
Government Regulation	**Anti-Trust Action** - Breaking up of monopolistic big tech companies
Government Regulation	**Taxes/Fines** - Taxing or fining social media companies for their use of personal user data
Gov or platform regulation	**Targeted Advertising** - Limiting or banning micro-targeted advertising
	8. GENERATIVE AI
Generative AI	**AI Chatbots**- The use of generative AI chatbots to reduce belief in conspiracy theories or misinformation
Generative AI	**Gen AI Content** - The use of generative AI to generate rebuttals to misinformation or create educational initiatives
Generative AI	**Deepfakes** - Prohibit usage of AI or manipulated content to misrepresent the speech or actions of public figures
Generative AI	**AI in Ads** - Prohibit usage of AI or manipulated content in political or targeted ads
Generative AI	**AI Disclosure** - Require clear disclosures on any ad that uses AI-generated images, video, or audio

- **Implementer:** Which organization(s) implement the intervention (e.g., social media platforms or the federal government).
- **Policy Changes:** Whether the intervention requires a new organizational policy, law, or regulation, and what those changes entail. This is often dictated by the implementer.
- **Targeted Phase:** Which part of the misinformation pipeline is the primary target of the intervention, such as the creation, spread, or belief.
- **Targeted Content:** The type of content being addressed (e.g., political misinformation, health misinformation).
- **Information Changes:** Whether the intervention requires design or algorithmic changes to the platforms, such as by modifying, removing, or tagging misinformation.

For example, the operationalized intervention *redirection* is typically implemented by platforms and focuses on mitigating the spread of misinformation, rather than its creation or belief, by reducing or eliminating its amplification to other users. Platforms determine the specific news contexts targeted by their policies, and redirection efforts often require updated platform policies specifying what is targeted, how it is targeted, and why. For example, platforms could clarify that authoritative public health sources, like the CDC or WHO, are prominently featured in search results while downranking unverified sources. Additionally, implementing redirection requires informational changes, such as modifying the information shown to users or the order in which it is presented.

These objective characteristics likely affect users' perceptions of the interventions, which in turn influence public policy support. Transparency is a particularly important factor in public policy support [118, 143]. A well-defined and transparent redirection policy, along with clearly marked informational changes, can shape an individual's perception of the fairness or intrusiveness of the intervention.

6.2.3 User Perceptions

User perception refers to how individuals evaluate the different characteristics of interventions and includes factors such as perceived effectiveness, fairness, intrusiveness, transparency, and problem awareness [118]. These perceptions influence whether people believe a problem even requires an intervention, whether individuals think the intervention will work, and whether they trust the proposed implementer to enforce the new policies fairly. The objective characteristics of the interventions affect user perceptions and are described in the following subsections. In addition to these characteristics, educational and messaging efforts more generally could also improve public perceptions. For example, a meta-analysis of public policy interventions demonstrated that effectively communicating a potential policy's effectiveness increased overall support by approximately 4% [243].

Implementer

Some individuals may trust certain implementers more than others. For example, partisanship is associated with lower levels of support for government regulation [295], and this finding was replicated in Chapter 5 when we observed that strong Republicans were more supportive of interventions implemented by platforms than of the same interventions implemented by governments.

Policy Changes

Transparent policy changes can enhance trust and, consequently, support for misinformation interventions [255]. Platforms should consider publishing examples of violating content and moderation guidelines [117]. For example, Google publishes and updates "Search Quality Rater Guidelines"[1], which details their entire page quality process as well as how their raters are trained to evaluate pages. However, trade-offs are involved, and more opaque policies may be less vulnerable to malicious actors attempting to exploit the platform's rules [143].

Targeted Phase

In Chapter 1, the misinformation pipeline is defined and addressed. The lifecycle of misinformation content on social media has three main phases: **creation, spread,** and **belief**. The creation aspect can refer to either the *network creation* or *content creation*. The spread, or distribution, of the content includes a direct *sharing* component as well as an algorithmic *amplification* component. Finally, the belief component can refer to the *verification* of content after the fact or the *prevention* of belief in misinformation. Chapter 5 and prior literature show that individuals prefer informational interventions over restrictive ones [87, 123]. This preference may lead to greater support for interventions aimed at targeting belief in misinformation rather than controlling content creation or distribution.

Targeted Content

Research indicates that public policy support often depends on problem awareness [118]. For example, prior work shows a higher level of support for anti-smoking initiatives compared to other similar interventions aimed at addressing alcohol consumption or diet. This difference can be partially explained by the public's awareness of the negative consequences associated with smoking [87]. Similarly, we observed in Chapter 4 that individuals indicate they are more willing to engage in social corrections for serious content. Therefore, people may be more willing to accept interventions for content they consider particularly severe.

Information Changes

Informational changes associated with interventions can also be implemented in explainable and transparent ways, increasing support levels [255]. Additionally, improved tool usability and platform reporting support have been shown to both increase users' willingness to engage with reporting systems as well as their perceived effectiveness [321].

[1] `https://services.google.com/fh/files/misc/hsw-sqrg.pdf` [Accessed 04-30-2025]

6.2.4 Main Criteria

Each operationalized intervention is assessed based on five primary characteristics: effectiveness, acceptance, effort level, cost, and political feasibility. These criteria were selected by reviewing previous systematic review articles and policy analyses.

In the public policy domain, Rochefort's comparative policy analysis of various social media regulations evaluates these policies using four measures: effectiveness, administrative difficulty, cost, and political acceptability [248]. Administrative burden, in terms of cost or implementation effort, is often applied in policy analysis [92, 168]. In Blair et al.'s review of misinformation interventions, the authors emphasize that, in addition to effectiveness, policymakers and practitioners designing interventions should also consider feasibility, scalability, and durability [46]. Feasibility refers to how easily an intervention can be implemented, while scalability addresses how simple it is to expand it for a larger audience. Durability, an aspect of effectiveness, reflects the longevity of an intervention's impact. It is often studied in the academic literature, particularly regarding the most commonly studied intervention types such as media literacy [121, 135], inoculation [186], and fact-checking [234]. Finally, Kozyreva et al.'s review of individual interventions discusses the ease of implementation and potential scalability in their conceptual overview of intervention types [167].

The five main criteria are defined as follows:

- **Effectiveness:** The extent to which an intervention reduces the creation, spread, or belief in misinformation under different circumstances, such as for some users, certain platforms, or particular types of content.
- **Acceptance:** The degree of user support for an intervention, which may be influenced by factors such as its perceived effectiveness, fairness, intrusiveness, transparency, and the level of trust in the implementer.
- **Effort level:** The level of effort required to implement and maintain the intervention for the users, platforms, or government entities involved.
- **Cost:** The financial burden on the users, platforms, or government entities involved in implementing and maintaining the intervention.
- **Political feasibility:** The likelihood of implementing an intervention, incorporating all previous factors as well as external factors, such as stakeholder perspectives, regulatory and legal constraints, and support from relevant officials or organizations.

Each intervention's objective characteristics and user perceptions directly influence the five key evaluative criteria used throughout this chapter. For example, user perceptions impact both **acceptance** and the degree of policy buy-in, which can, in turn, increase policy compliance and **effectiveness** [118]. Similarly, interventions that require substantial policy or informational changes may increase administrative burden, raising both **effort** and **cost** [248]. **Political feasibility** incorporates these factors as well as several external elements, such as governmental or platform structures, legal considerations, the influence of other stakeholders, and the broader political climate [248]. For example, if the administrative burden of a policy (high cost and implementation effort) outweighs the potential benefits (effectiveness of the intervention), such policies may not be implemented even if they are both effective and acceptable to the public [168].

An example of the impact of political feasibility is Meta's recent removal of its fact-checking program [149]. Despite evidence of its effectiveness [115, 173] and broad user acceptance by most users across the political spectrum [239], external pressures may have influenced this decision [319]. These external factors may have included fear of potential government regulation or allegations of perceived anti-conservative bias [319]. While fact-checking is generally popular, including among a majority of Republicans, support is stronger among Democrats and Independents than among Republicans [112, 239]. Notably, recent work also finds that Republicans perceive professional fact-checkers as a more legitimate means of doing content moderation compared with regular social media users [188], and relying solely on community notes is actually extremely unpopular across the political spectrum [239].

6.3 Data and Methods

A survey was designed to gather expert opinions on intervention effectiveness, acceptance, effort level, cost, and political feasibility. The survey was promoted to participants and invitees for the *Disinformation and AI Summit* at Carnegie Mellon University, which took place from January 22 to 24, 2025. Data was collected between January 23 and March 24, 2025.

6.3.1 Survey Design

The survey was conducted using Qualtrics and consisted of four main sections: Introduction and Consent, General Category Questions, Intervention Questions, and Demographics and Survey Feedback.

Introduction and Consent

Participants were informed that the survey focused on interventions to counter misinformation. They were also told that the survey would take approximately 10 minutes, that there were no risks or compensation associated with the survey, that they could exit the survey at any time, and that all responses would be anonymized in any analysis of the results. They were asked to provide their informed consent to participate before proceeding with the survey.

General Categories

In this section, participants were asked for their professional opinions on the eight general categories of misinformation interventions defined in this dissertation (Table 1.5). The only difference is that the **Other** category was redefined as a **Generative AI** category. More specifically, participants were asked two primary questions.

1. To what extent do you agree that the following categories of misinformation interventions **are effective** at combatting misinformation if adopted widely by social media companies

or institutions? *[Strongly disagree, Disagree, Tend to disagree, Neither agree nor disagree, Tend to agree, Agree, Strongly agree]*

2. To what extent do you agree that the following categories of misinformation interventions would be **acceptable** to social media users or citizens if adopted widely by social media companies or institutions? *[Strongly disagree, Disagree, Tend to disagree, Neither agree nor disagree, Tend to agree, Agree, Strongly agree]*

For each question, participants were instructed to consider, compare, and rate all eight categories simultaneously using a 1-7 Likert Scale. They were provided with the following definitions for each general category, summarized from Appendix A:

- **Content Distribution:** Refers to how content is distributed on social media. This includes interventions such as redirection, accuracy prompts, friction, platform alterations, and advertising policy
- **Content Moderation:** Refers to how content is shown or not shown on social media. This includes interventions such as fact-checking, debunking, misinformation detection, and algorithmic content moderation
- **Account Moderation:** Refers to how accounts are moderated on social media. This includes suspending, banning, shadow banning, or demonetizing user accounts.
- **Content Labeling:** Refers to misinformation disclosure involving labels of any kind. This includes fact-checking labels, crowdsourcing labels like community notes, warning labels, source credibility labels, and context labels
- **User-based Measures:** Refers to measures that involve other users seeing misinformation and how they respond to it. This includes users reporting or blocking other users or their posts, using social corrections, or updating social norms.
- **Media Literacy and Education:** Refers to any educational or training effort meant to increase the public's civic reasoning and critical thinking skills. Includes fake news games and other inoculation methods
- **Institutional Measures:** Refers to how governments and other public or civic institutions can help manage the spread of misinformation. This includes interventions such as investing in or promoting local news, allowing researchers access to data, and government regulation.
- **Generative AI:** The usage of generative AI to detect misinformation, generate rebuttals, create educational resources, or dialoguing with those who believe misinformation.

Specific Interventions

In this section, participants were asked to provide their professional opinions on the specific, operationalized interventions (see Tables 6.1 and 6.2). Each participant randomly viewed 12 of the 40 interventions described in those tables. They were informed that they would be asked about each potential intervention's effectiveness, user acceptance, effort level, cost, and political feasibility. They rated each intervention across these five factors using a 1-5 Likert scale, similar to the one implemented in Chapter 5.

- Would the proposed policies be **effective** or **ineffective** in combatting misinformation if adopted widely by social media companies or institutions? *[Very ineffective, Somewhat ineffective, Neither effective nor ineffective, Somewhat effective, Very effective]*
- Would the proposed policies be **acceptable** or **unacceptable** to users or citizens if adopted widely by social media companies or institutions? *[Very unacceptable, Somewhat unacceptable, Neither acceptable nor unacceptable, Somewhat acceptable, Very acceptable]*
- Would the proposed policies require a **high or low amount of time and effort** for companies or institutions to implement? *[Very low effort, Somewhat low effort, Neither high nor low effort, Somewhat high effort, Very high effort]*
- Would the proposed policies require a **high or low amount of financial investment** for companies or institutions to implement? *[Very low cost, Somewhat low cost, Neither high nor low cost, Somewhat high cost, Very high cost]*
- Would the proposed policies be **politically feasible or infeasible** for companies or institutions to implement? *[Very infeasible, Somewhat infeasible, Neither feasible nor infeasible, Somewhat feasible, Very feasible]*

Demographics and Survey Feedback

Participants were given optional demographic questions to answer, including academic rank (PhD student, postdoc, faculty, industry, or other), age category, gender, and primary area of expertise (Psychology, Communication and Media Studies, Political Science, Journalism, Computational Social Sciences, Sociology, Computer Science, or Other). Finally, participants had the opportunity to share their thoughts and feedback.

6.3.2 Participant Demographics

Thirty-nine participants completed the survey. A high-level summary of participant demographics is provided below:

- **Gender:** 16 women (41.0%), 22 men (56.4%), 1 other/prefer not to answer
- **Age:** 18-34 (15 people, 38.5%), 35-44 (12 people, 30.8%), 45+ (10 people, 25.6%), 2 Prefer not to answer
- **Academic Rank:** 21 faculty (53.8%), 12 PhD students (30.8%), 2 Industry (5.1%), 1 Postdoc (2.6%), 3 Other (7.7%).
- **Primary Discipline:**
  - Computational social sciences: 17 (43.6%)
  - Computer science: 10 (25.6%)
  - Sociology: 4 (10.3%)
  - Communications and media studies: 4 (10.3%)
  - Political science: 2 (5.1%)
  - Other: 2 (5.1%)

6.3.3 Analysis Methods

This chapter analyzes the values from the expert survey and summarizes the results from the previous chapters and the literature review, if applicable. If available, meta-analyses and systematic review articles on the effectiveness of specific interventions will be included in the discussion. Any available sources for the other four metrics, such as public opinion polls or policy analyses, will also be addressed.

In general, an ideal intervention would have high acceptance, effectiveness, and political feasibility while requiring low effort and cost. Therefore, to better compare interventions, effort level and cost scores were inverted so that 1 indicates very high effort or cost and 5 indicates very low effort and cost. This was done to ensure that high Likert scores for all five factors represented the ideal outcome. The average Likert scores across the five metrics will be calculated for each intervention and used to compare and rank them.

6.3.4 Measures

Effectiveness for interventions by general category: These responses are coded from 1 to 7 (strongly disagree to strongly agree).

Acceptance for interventions by general category: These responses are coded from 1 to 7 (strongly disagree to strongly agree).

Effectiveness for specific intervention(s): These responses are coded from 1 to 5 (least to most effective)

Acceptance for specific intervention(s): These responses are coded from 1 to 5 (least to most acceptable).

Effort level for specific intervention(s): These responses are coded from 1 to 5 (very high to very low effort).

Cost for specific intervention(s): These responses are coded from 1 to 5 (very high to very low cost).

Political feasibility for specific intervention(s): These responses are coded from 1 to 5 (very infeasible to very feasible).

6.4 Results and Discussion

First, the effectiveness and acceptance ratings for each general category of misinformation interventions are assessed. Next, each category of interventions is analyzed in detail to evaluate the operationalized interventions within each category. The categories are examined in the order they are presented in Appendix A. The objective characteristics of the interventions are reviewed, and the results of the expert survey analysis on the five main factors are analyzed. The results section concludes with a summary of the top interventions for each of the five main metrics.

Comparison of General Categories

Figure 6.1 and Table 6.3 show the fraction of participants who believed each category of countermeasures to be effective or acceptable to the public, sorted by the proportion who believe it is effective. Content Moderation, Account Moderation, and Media Literacy / Education are regarded as the most effective, with over 70% of respondents agreeing they would be effective. However, moderation, especially Account Moderation, is viewed by experts as one of the least acceptable to the public.

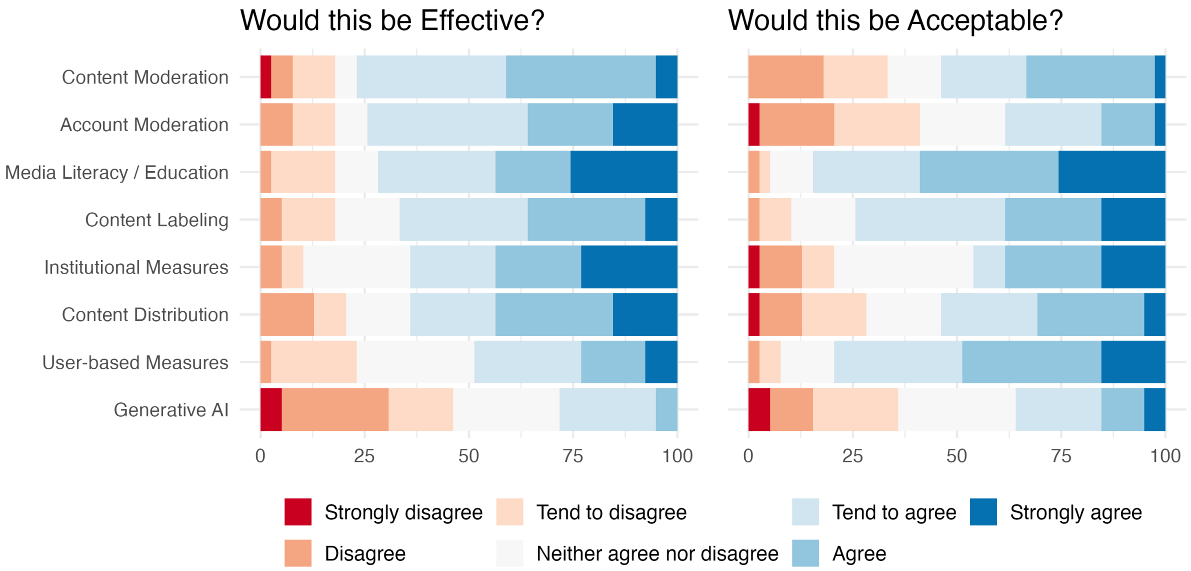


Figure 6.1: Stacked bar plot representing the percentage of responses on the effectiveness and acceptance of the general categories of interventions.

Figure 6.2 illustrates the average Likert scores and 95% confidence intervals for the effectiveness and acceptance ratings, sorted by the highest average effectiveness score. In this figure, we observe that Media Literacy / Education and Institutional Measures rise in the effectiveness rankings, bolstered by the proportion of participants who "strongly agree" regarding their effectiveness rather than more weakly agreeing. Content and Account Moderation remain at high levels. This figure more clearly highlights the differences in effectiveness and acceptance across various intervention categories. Account Moderation and User-based Measures exhibit the largest gap between effectiveness and acceptance, with effectiveness exceeding acceptance in Account Moderation, and acceptance exceeding effectiveness in User-based Measures.

These results align with a similar survey of experts, which found that they believed media literacy, labeling of false content, and fact-checking to be the most effective interventions, each receiving around 70% approval. Meanwhile, accuracy prompts (a form of content distribution), source labeling, and inoculation ranked lower on the list. These were the only interventions included in this survey in a similarly structured question [16]. This survey also asked experts if certain types of actions *should* be taken, with platform design changes, algorithmic changes, and general content moderation as the most popular (over 75% agree), and penalizing misin-

	Effectiveness			Acceptance		
Category	**Agree**	**Disagree**	**Neither**	**Agree**	**Disagree**	**Neither**
Content Moderation	76.9	17.9	5.13	53.8	33.3	12.8
Account Moderation	74.4	17.9	7.69	38.5	41.0	20.5
Media Literacy / Education	71.8	17.9	10.3	84.6	5.13	10.3
Content Labeling	66.7	17.9	15.4	74.4	10.3	15.4
Institutional Measures	64.1	10.3	25.6	46.2	20.5	33.3
Content Distribution	64.1	20.5	15.4	53.8	28.2	17.9
User-based Measures	48.7	23.1	28.2	79.5	7.69	12.8
Generative AI	28.2	46.2	25.6	35.9	35.9	28.2

Table 6.3: The percentage of respondents who agree or disagree that each general category of interventions is effective or acceptable to the public.

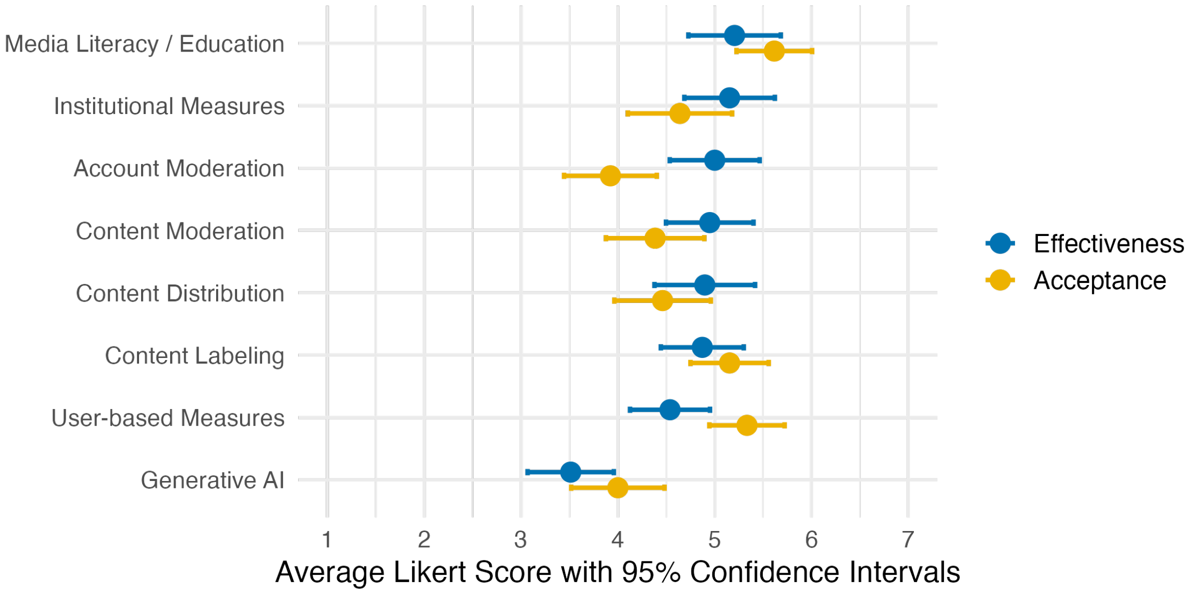


Figure 6.2: Average effectiveness and acceptance Likert scores per general intervention category.

formation sharing and shadow banning (a form of account moderation) as the least popular. In fact, shadow banning was the only suggested intervention where more experts disagreed with its implementation than agreed [16]. Similarly, Blair et al. found in a survey of experts that they recommended implementing educational and institutional interventions (such as media literacy, platform alterations, and journalist training) over other types of interventions [46]. However, the authors note that these interventions tend to be among the least well-studied or have mixed evidence on actual effectiveness [46].

6.4.1 Content Distribution

This section presents a comparative analysis of the operationalized content distribution interventions (refer to Table 6.1 or Appendix A.1 for detailed definitions).

Objective Characteristics

Table 6.4 shows the objective characteristics of the eight operationalized content distribution interventions, listed in the order they were initially presented in Table 6.1. This table illustrates the typical **implementer** of each intervention, the **targeted phase** of the misinformation pipeline, and the **information changes** often required by the intervention. Two objective features (**policy changes** and **targeted content**) are not included in the table as they depend heavily on the implementing organization's policies and the overall implementation of the intervention.

Table 6.4: Objective characteristics of Content Distribution interventions.

Intervention	**Implementer**	**Targeted Phase**	**Info Changes**
Redirection	Platforms	Spread (sharing)	Modifies, Removes
Accuracy Prompts	Platforms	Spread (sharing)	Adds
Friction	Platforms	Spread (sharing)	Adds
Platform Alterations	Platforms	Spread (both)	Modifies
Ban Political Ads	Platforms, Govts	Spread (sharing)	Removes
Fact-Check Ads	Platforms, Govts, Insts	Spread (amplification) Belief (verification)	Tags, Removes
Limit Forwarding	Platforms	Spread (amplification)	Removes
Limit Resharing	Platforms	Spread (amplification)	Removes

The implementer column outlines the typical implementer of each intervention. Sometimes, these interventions could be carried out by one or more implementers. For example, banning political advertising on social media platforms may originate from the platforms themselves or be mandated by the government, which would still require the platforms to implement it. Fact-checking advertisements might involve third-party fact-checking organizations. When platforms adopt many of these content distribution interventions, such as redirection, friction, advertising policies, and restrictions on resharing or forwarding, they are usually detailed in their platform policies (see Section 1.5), though to varying degrees of detail.

Most content distribution interventions target the spread of misinformation. The only exception is fact-checking advertisements, which, depending on implementation, may also address the belief in the misinformation by allowing the content to remain on the platform but with labels, for example. Several content distribution interventions, such as redirecting users to official content or reminding them about accuracy, specifically target the initial sharing of misinformation by user accounts. Even the complete banning of political advertisements essentially addresses the initial sharing of content by advertising organizations. Other content distribution interventions, such as platform alterations like reducing the size of a post containing misinformation or capping the number of accounts one can reshare content to, limit the potential for amplification.

When considering the informational changes associated with implementing the interventions, both accuracy prompts and friction require adding interstitials or other content or UI changes. Redirection and platform alterations change how information is displayed to users. Banning political advertising removes content, while limiting forwarding or resharing also removes functionality.

Expert Survey Results

Table 6.5 summarizes the expert scores for the content distribution interventions, ranked by highest average score across all five criteria.

The two highest-ranked content distribution interventions involved limiting the excessive resharing or forwarding of messages. Experts found these interventions to be relatively effective, acceptable to users, low effort, low cost, and politically feasible. The literature on these interventions is limited; they are likely effective and relatively easy to implement and scale [6, 235]. However, more studies are needed on both their effectiveness and user acceptance.

Table 6.5: Expert scores for Content Distribution interventions. Mean values are shown with their corresponding standard deviations in parentheses.

Intervention	N	Effectiveness	Acceptance	Effort	Cost	Feasibility	Avg
Limit Resharing	9	3.78 (0.44)	3.00 (1.22)	4.22 (0.67)	4.56 (0.53)	4.00 (1.22)	3.91
Limit Forwarding	13	3.69 (0.63)	3.54 (1.05)	3.92 (1.04)	4.31 (0.85)	4.08 (0.86)	3.91
Friction	12	3.50 (1.00)	3.33 (1.07)	3.33 (1.15)	3.83 (1.47)	3.75 (0.97)	3.55
Accuracy Prompts	8	3.62 (0.92)	3.50 (1.07)	3.38 (1.06)	2.88 (1.13)	3.38 (1.51)	3.35
Platform Alterations	10	3.80 (1.32)	3.00 (0.94)	2.50 (0.97)	3.10 (1.37)	3.60 (1.07)	3.20
Redirection	10	3.10 (0.99)	3.10 (1.10)	3.10 (1.10)	3.90 (0.88)	2.60 (1.35)	3.16
Ban Political Ads	15	3.00 (0.93)	3.21 (1.37)	3.60 (1.30)	3.27 (1.49)	2.20 (1.32)	3.06
Fact-Check Ads	9	3.44 (0.73)	3.75 (0.71)	2.00 (0.93)	2.00 (0.93)	3.33 (1.00)	2.91

Friction and accuracy prompts were ranked as the next highest. While the literature suggests that friction may be effective, more work is needed [29, 45, 262]. Friction was analyzed in Chapter 5 and was the least popular intervention, with an average Likert score of 2.6, despite its frequent usage by platforms. Additionally, it received low ratings for perceived effectiveness and fairness and was seen as highly intrusive. This result may have been due to the phrasing of the intervention in the survey. Further user acceptance research should be conducted on friction to determine the best implementation strategy.

Meanwhile, accuracy prompts are widely studied in the literature. Although effect sizes vary, they are generally considered effective and relatively easy to implement and scale [46, 227].

A meta-analysis found that accuracy prompts can reduce the sharing of false information by approximately 10% [227]. However, user acceptance of accuracy prompts has not been studied as extensively, although it may be similar to the acceptance of related interventions such as friction.

Platform alterations and redirection received generally positive evaluations from experts. The literature suggests that both are generally effective, have moderate acceptance levels, and are relatively easy to implement and scale, although more studies are needed [45, 47, 160]. Redirection through related articles and similar methods has typically been shown to be effective according to the literature [45, 47]. In terms of user acceptance, one prior study found that over half of participants supported the use of related articles combined with a reduction in the size of misinformation, while less than a quarter opposed it [160].

Finally, advertising policy ranked the lowest among the content distribution interventions. Fact-checking ads, in particular, was viewed as particularly effortful and costly, likely due to the need to engage with fact-checking organizations. However, the American public generally supports banning or limiting political ads on social media platforms [23]. Similarly, the public generally supports fact-checking ads, and fact-checking has consistently been found to be effective [9, 302]. Fact-checking ads ranked among the most popular interventions tested in Chapter 5, with an average Likert score of 4 (indicating "high" support). They also received high scores for perceived effectiveness and fairness and were rated at a medium level of intrusiveness. However, more academic research is needed to better understand the effectiveness of advertising policy.

6.4.2 Content Moderation

This section presents a comparative analysis of the operationalized content moderation interventions (refer to Table 6.1 or Appendix A.2 for detailed definitions).

Objective Characteristics

Table 6.6 shows the objective characteristics of the five operationalized content moderation interventions, listed in the order they were initially presented in Table 6.1.

Table 6.6: Objective characteristics of Content Moderation interventions.

Intervention	Implementer	Targeted Phase	Info Changes
Content Removal	Platforms, Govts	Spread (amplification)	Removes
Downranking	Platforms	Spread (amplification)	Removes
Algorithmic Changes	Platforms	Spread (amplification)	Adds, Removes
Virality Circuit Breakers	Platforms	Spread (amplification)	Removes
User Control	Platforms, Users	Spread (amplification)	Adds, Removes

Content moderation interventions are similar to those for content distribution in several ways. These interventions are usually implemented almost exclusively by platforms. However, governments may sometimes impose specific restrictions on content that must be removed, such as illegal content. Allowing users greater control over their news feeds requires implementation by the platforms, as well as engagement from the users. Content moderation interventions also typically target the spread of misinformation, focusing on limiting or removing the algorithmic amplification of the content rather than addressing the initial sharing component of the spread.

When considering the informational changes associated with implementing these interventions, most content moderation strategies involve some level of content removal or downranking after the content has already been shared. However, algorithmic upranking of credible content and allowing users more control over their feeds may also require adding content through design or UI changes.

Expert Survey Results

Table 6.7 summarizes the expert scores for the content moderation interventions, ranked by highest average score across all five criteria.

User control is the top-ranked content moderation strategy because of its high perceived user acceptance and potential feasibility. Considering that in Chapter 5 we found most people preferred less intrusive and restrictive interventions, these findings align with the expert opinions. Previous studies indicate that users are in favor of both personal controls over moderation and platform moderation regarding harmful content, such as hate speech or violent material. However, when given a choice, they tend to prefer personal moderation tools over a top-down, platform-based approach due to concerns about free speech [142]. However, there is limited research on the effectiveness of users controlling their news feeds.

Table 6.7: Expert scores for Content Moderation interventions. Mean values are shown with their corresponding standard deviations in parentheses.

Intervention	N	Effectiveness	Acceptance	Effort	Cost	Feasibility	Avg
User Control	14	3.00 (1.18)	4.36 (0.84)	2.31 (0.95)	3.38 (1.39)	4.21 (1.12)	3.45
Algorithmic Changes	11	4.09 (0.54)	3.55 (1.04)	2.55 (0.93)	3.00 (1.26)	3.64 (1.03)	3.36
Content Removal	9	3.89 (0.93)	3.33 (0.87)	2.44 (1.24)	3.78 (0.97)	3.11 (1.17)	3.31
Downranking	13	3.92 (0.76)	3.23 (1.17)	3.08 (1.04)	3.15 (1.28)	3.15 (1.46)	3.31
Virality Circuit Breakers	12	3.58 (1.38)	3.08 (0.90)	2.92 (1.00)	3.00 (1.28)	3.58 (1.08)	3.23

Algorithmic upranking of credible content was ranked the next highest by experts, beating out the removal of content verified to contain misinformation and algorithmic downranking. The virality circuit breakers intervention, which involves temporarily stopping the spread of fast-spreading unverified content, was at the bottom of the ranking [6]. These circuit breakers, while typically only temporary, involve a complete block-out of that piece of content on the platform while the content is being verified, which can affect the ability of user accounts to monetize their content.

While changes to recommendation algorithms are often found to be effective [62], these expert results are in line with our previous findings from Chapter 5. Content removal and downranking were among the least popular interventions in our survey, both receiving moderate support levels of around 3.6 on a Likert scale, though this support was higher if the implementer was a platform rather than the government. While participants perceived content removal to be relatively effective (3.7), it was also seen as one of the most intrusive interventions in the survey (3.5). Algorithmic downranking was perceived as less intrusive (3.2) but also less effective (3.1).

6.4.3 Account Moderation

This section presents a comparative analysis of the operationalized account moderation interventions (refer to Table 6.1 or Appendix A.3 for detailed definitions).

Objective Characteristics

Table 6.8 shows the objective characteristics of the four operationalized account moderation interventions, listed in the order they were initially presented in Table 6.1.

Table 6.8: Objective characteristics of Account Moderation interventions.

Intervention	**Implementer**	**Targeted Phase**	**Info Changes**
Account Suspensions	Platforms	Creation (network)	Removes
Deplatforming	Platforms	Creation (network)	Removes
Shadow Banning	Platforms	Spread (amplification)	Removes
Demonetization	Platforms	Creation (content), Spread (amplification)	Removes

Account moderation interventions are similar to those for content moderation in several ways. These interventions are typically carried out almost exclusively by platforms and involve some level of account or content removal. However, by moderating users instead of content, account moderation strategies generally aim to combat the network or account creation phase of the misinformation pipeline rather than its spread or belief. Shadow banning is an exception, as it allows users to remain on the platform while limiting or not sharing their posted content with others [308]. It is the application of algorithmic downranking specifically for user accounts. Similarly, demonetization of a user account may influence both that user's willingness to create content and the type of content they create, and it can sometimes the spread of a user's content [190].

Expert Survey Results

Table 6.9 summarizes the expert scores for the account moderation interventions, ranked by highest average score across all five criteria.

Among the account moderation strategies, demonetization received the highest ratings from misinformation researchers. The other interventions scored relatively low in potential user acceptance, which aligns with both previous expert surveys [16] and our findings in Chapter 5. The second least popular intervention in our public opinion survey was permanently banning users. Although prior work demonstrates the effectiveness of account suspensions and deplatforming for particularly problematic individuals [191, 241, 280], considerably less research has been conducted on demonetization or shadow banning [144].

Table 6.9: Expert scores for Account Moderation interventions. Mean values are shown with their corresponding standard deviations in parentheses.

Intervention	N	Effectiveness	Acceptance	Effort	Cost	Feasibility	Avg
Demonetization	12	4.00 (0.60)	3.75 (1.06)	2.92 (1.24)	2.92 (1.51)	3.58 (1.31)	3.43
Shadow Banning	9	3.33 (1.32)	2.89 (1.05)	3.38 (1.51)	3.50 (1.69)	3.44 (1.24)	3.31
Account Suspensions	15	3.53 (0.99)	2.86 (1.29)	2.40 (0.83)	3.47 (1.13)	2.80 (1.52)	3.01
Deplatforming	12	3.83 (0.83)	2.92 (1.00)	2.33 (1.07)	3.25 (1.36)	2.58 (1.31)	2.98

While likely effective, account moderation may be one of the least popular countermeasures categories among social media users, which could affect the likelihood of implementation. However, demonetization scored relatively well compared to other interventions in this category, especially regarding acceptance and feasibility. This may be due to its less punitive and restrictive nature than suspensions or deplatforming, and its more transparent nature compared to shadow banning. Often, users are unaware of a shadow ban, but demonetization would have a more direct impact.

6.4.4 Content Labeling

This section presents a comparative analysis of the operationalized content labeling interventions (refer to Table 6.1 or Appendix A.4 for detailed definitions).

Objective Characteristics

Table 6.10 shows the objective characteristics of the five operationalized content labeling interventions, listed in the order they were initially presented in Table 6.1.

Table 6.10: Objective characteristics of Content Labeling interventions.

Intervention	Implementer	Targeted Phase	Info Changes
Crowdsourcing	Platforms, Users	Belief (verification)	Adds, Tags
Click-through Warnings	Platforms	Belief (verification, prevention)	Adds, Modifies
Source Labeling	Platforms	Belief (verification)	Adds, Tags
Government Labels	Platforms	Belief (verification)	Adds, Tags
Fact-Check Labels	Platforms, Insts	Belief (verification)	Adds, Tags

Content labeling interventions resemble those for content moderation in several ways, as they are a form of content moderation typically conducted after the misinformation has already spread and potentially peaked. These interventions are implemented almost exclusively by platforms. However, institutions may be involved in creating fact-checking labels or links, and users contribute to crowdsourcing context labels that are added to existing posts.

These interventions aim to address belief in misinformation by warning users about the misleading content and attempting to prevent such beliefs, or by using verification techniques to correct existing false content and associated misconceptions. When considering the informational changes associated with implementing these interventions, most labeling strategies involve some level of additional, modified, or tagged information on content after it has already been shared.

Expert Survey Results

Table 6.11 summarizes the expert scores for the content labeling interventions, ranked by highest average score across all five criteria.

Overall, content labeling interventions received high scores, particularly in the user acceptance category. These findings are consistent with our previous results from Chapter 5, in which we found that the two content labeling interventions included in our survey, including fact-check labels, ranked as the top two most supported interventions overall in that survey.

Table 6.11: Expert scores for Content Labeling interventions. Mean values are shown with their corresponding standard deviations in parentheses.

Intervention	N	Effectiveness	Acceptance	Effort	Cost	Feasibility	Avg
Government Labels	5	3.80 (1.30)	4.00 (0.00)	4.00 (0.00)	4.40 (0.55)	4.00 (0.00)	4.04
Crowdsourcing	14	3.21 (0.80)	4.08 (0.64)	3.07 (1.27)	4.14 (1.17)	4.29 (0.61)	3.76
Source Labeling	9	3.67 (0.87)	3.67 (0.87)	2.33 (1.00)	3.56 (1.24)	3.11 (1.17)	3.27
Fact-check Labels	15	3.87 (0.74)	3.80 (0.94)	2.21 (1.12)	2.79 (1.12)	3.53 (1.06)	3.24
Click-through Warnings	12	3.75 (0.97)	3.17 (1.03)	2.42 (1.08)	3.17 (1.40)	3.42 (1.08)	3.18

Labeling government-owned or sponsored media accounts received the highest scores from experts in this survey, followed by crowdsourcing context labels. Previous research has demonstrated that government labels can effectively reduce engagement with Russian-sponsored election misinformation if implemented properly by the platforms [208]. The evidence so far on the effectiveness of crowdsourced Community Notes programs is more mixed. Studies have shown that they have increased the amount of content labeled, but they may not have a rapid enough response time to tackle fast-spreading misinformation [71, 90, 148]. The evidence for context labels, more broadly, has also been found to be mixed, though further research is necessary [46]. Among experts in this survey, it ranks the lowest in effectiveness compared with other content labeling interventions.

Source credibility labels and fact-check labels ranked as the next highest scoring interventions. In a large-scale review, Blair et al. found that credibility labels were relatively effective, although not as effective as inoculation or prebunking [45]. Credibility labels that reference expert fact-checkers or organizations and clearly indicate whether the content is true or false (rather than just "disputed") tend to be more effective than other types of labels [45, 160, 196].

Finally, click-through warning labels ranked the lowest, partly due to their low user acceptance score, which is likely related to their perceived intrusiveness on the user experience and relative inconvenience. While they are likely generally effective at reducing misinformation belief [262], more work on both effectiveness and user acceptance is needed. Because the perceived intrusiveness of this intervention is likely high, addressing user acceptance may present challenges.

6.4.5 User-based Measures

This section presents a comparative analysis of the operationalized user-based interventions (refer to Table 6.2 or Appendix A.5 for detailed definitions).

Objective Characteristics

Table 6.12 shows the objective characteristics of the three operationalized user-based interventions, listed in the order they were initially presented in Table 6.2.

Table 6.12: Objective characteristics of User-based interventions.

Intervention	Implementer	Targeted Phase	Info Changes
Reporting	Platforms, Users	Spread (amplification)	Adds, Modifies
Social Norms	Platforms, Users	Belief (verification)	Adds, Modifies
Alter Platform Metrics	Platforms	Creation (content), Spread (both)	Modifies

Platforms can implement user-based interventions by encouraging active user participation. User reporting tools enable users to notify the platform about potentially problematic content. If the content violates platform policies, it may be removed or its distribution may be restricted, addressing the potential for further amplification of the problematic content. The social norms intervention was described in Table 6.2 as encouraging self or social corrections after unintentionally spreading misinformation. This intervention primarily addresses the belief verification phase of the misinformation pipeline, though it could also affect additional content sharing. Similarly, altering platform metrics to reward accuracy instead of engagement can influence the likelihood that misinformation is created and shared on the platform, as well as the likelihood that others will engage with it or spread it further. Platforms can encourage engagement with user-based measures by adding or modifying existing tools on their sites, such as improving reporting functionality and transparency or altering content engagement metrics.

Expert Survey Results

Table 6.13 summarizes the expert scores for the user-based interventions, ranked by highest average score across all five criteria.

Table 6.13: Expert scores for User-based Measures. Mean values are shown with their corresponding standard deviations in parentheses.

Intervention	N	Effectiveness	Acceptance	Effort	Cost	Feasibility	Avg
Reporting	10	3.80 (0.79)	4.00 (0.94)	2.40 (1.07)	3.00 (1.49)	4.40 (0.52)	3.52
Alter Platform Metrics	12	3.92 (1.00)	3.75 (1.36)	2.50 (1.09)	3.00 (1.41)	4.08 (0.79)	3.45
Social Norms	13	3.62 (0.96)	3.83 (0.94)	2.67 (1.37)	3.33 (1.23)	3.62 (0.77)	3.41

The user-based interventions received high scores and were one of the highest-ranking categories overall. Informational interventions and those that give users control and agency were common themes we noted for the most supported interventions in Chapter 5. Improved reporting functionality [321] and the usage of social norms to encourage pro-social behavior have been shown to be effective in prior work [46, 98, 110]. Platforms should solicit feedback and utilize user experience research techniques to improve existing tool usability and design improved systems in the future [117]. While this recommendation is valid for all interventions, especially newer ones where less research has been conducted, it is especially important for user-based interventions, as they require active user participation and satisfaction to be effective.

6.4.6 Media Literacy and Education

This section presents a comparative analysis of the operationalized media literacy interventions (refer to Table 6.2 or Appendix A.6 for detailed definitions).

Objective Characteristics

Table 6.14 shows the objective characteristics of the two operationalized media literacy interventions, listed in the order they were initially presented in Table 6.2.

Table 6.14: Objective characteristics of Media Literacy interventions.

Intervention	Implementer	Targeted Phase	Info Changes
Digital Media Literacy	Platforms, Govts, Insts	Belief (prevention)	Adds
Inoculation	Platforms, Govts, Insts	Belief (prevention)	Adds

Media literacy and educational efforts can be created and implemented by various institutions, including companies, governments, schools, and civic organizations. These initiatives typically aim to reduce belief in misinformation by improving people's competencies on social media. Organizations that adopt media literacy initiatives often integrate these efforts into their formal policies (for example, YouTube mentions their investment in these efforts on their platform policies page; see Section 1.5). If these fake news games or inoculation efforts are implemented directly on the platforms, then the platforms would need to add content to their sites, requiring software or other information changes.

Expert Survey Results

Table 6.15 summarizes the expert scores for the media literacy interventions, ranked by highest average score across all five criteria.

Table 6.15: Expert scores for Media Literacy and Education interventions. Mean values are shown with their corresponding standard deviations in parentheses.

Intervention	N	Effectiveness	Acceptance	Effort	Cost	Feasibility	Avg
Digital Media Literacy	9	3.78 (1.20)	4.11 (0.93)	2.00 (1.12)	2.00 (1.12)	3.78 (0.83)	3.13
Inoculation	10	3.40 (0.84)	3.90 (0.88)	2.10 (0.99)	2.50 (0.97)	3.60 (0.84)	3.10

Media literacy interventions had middling scores, dragged down by their perceived high effort levels and implementation costs. There is mixed evidence in the literature regarding the effectiveness of media literacy initiatives in improving discernment [21, 45, 121, 146]. Specifically, media literacy efforts have not been shown to be particularly durable over time without repeated sessions, in contrast to more standard countermeasures like debunking. However, the effects are likely greater in the Global South compared to the Global North [45, 121]. Furthermore, they are less scalable than other types of interventions [37], such as accuracy prompts [45]. However, our survey for Chapter 5 found that digital media literacy was one of the more popular interventions, with an average Likert support score of 4. It was also perceived as the fairest intervention proposed (average fairness score of 4.1).

Meanwhile, inoculation has generally proven effective in the short term, though fake news games and technique inoculation have more mixed results [46]. In direct comparisons, inoculation has been found to be more effective than credibility labels, yet less effective than debunking [45, 56].

6.4.7 Institutional Measures

This section presents a comparative analysis of the operationalized institutional interventions (refer to Table 6.2 or Appendix A.7 for detailed definitions).

Objective Characteristics

Table 6.16 shows the objective characteristics of the eight operationalized institutional interventions, listed in the order they were initially presented in Table 6.2.

Table 6.16: Objective characteristics of Institutional Measures interventions.

Intervention	Implementer	Targeted Aspect	Info Changes
Media Support	Platforms, Govts, Insts	Belief (prevention)	-
Journalism Support	Platforms, Govts, Insts	Belief (prevention)	-
Data Sharing	Platforms, Govts, Insts	-	Adds
Government Regulation	Govts	Spread (amplification)	Modifies, Removes
Privacy Legislation	Govts	-	-
Anti-Trust Action	Govts	-	-
Taxes / Fines	Govts	-	-
Targeted Advertising	Platforms, Govts	Creation (network), Spread (amplification)	Modifies, Removes

Institutional measures are similar to media literacy efforts in several ways. They can also be created and implemented by various institutions, such as companies, governments, schools, and civic organizations. Platforms can support local media and journalists and improve data transparency. However, the primary difference is that the impact of institutional interventions is often offline or more indirect in nature, not always specifically targeting any phase in the misinformation lifecycle on social media directly. While supporting local news and journalists can indirectly help prevent belief in misinformation by improving competencies, other interventions, such as data sharing, may increase researchers' knowledge and inform future intervention design without directly addressing the misinformation pipeline. Additionally, data sharing and transparency reports often require additional information to be posted on these social media sites, necessitating information changes on the platforms. Similarly, limiting or prohibiting targeted advertising may also require informational changes or removals.

There are several interventions that can only be implemented by governments, such as regulation on allowable content, privacy legislation, anti-trust action, and the levying of taxes or fines. Regulation of content on platforms targets the spread of misinformation similarly to other content moderation interventions by requiring platforms to modify or remove such content. Other government interventions address misinformation more indirectly by tackling privacy concerns or the monopolistic tendencies of social media companies. Depending on their implementation, these measures could affect the information on social media platforms and target multiple parts of the misinformation pipeline.

Expert Survey Results

Table 6.17 summarizes the expert scores for the institutional interventions, ranked by highest average score across all five criteria.

Institutional interventions vary significantly in their overall average scores and individual metric scores. The government-driven interventions all ranked in the bottom half due to their perceived high level of implementation effort, cost, and relatively low political feasibility. This is despite experts generally believing that the proposed measures would be effective or have popular support. Limiting or restricting targeted advertising and increased data sharing and transparency received the highest scores. The lack of data access and sharing poses a serious problem for the research community when investigating the effectiveness or user acceptance of platform-specific interventions [22, 52]. Improved data-sharing practices could, therefore, positively impact multiple types of interventions, including those in other categories.

Table 6.17: Expert scores for Institutional Measures. Mean values are shown with their corresponding standard deviations in parentheses.

Intervention	N	Effectiveness	Acceptance	Effort	Cost	Feasibility	Avg
Targeted Ads	13	3.38 (0.65)	3.85 (0.80)	3.33 (1.30)	3.08 (1.38)	3.38 (1.19)	3.41
Data Sharing	11	4.09 (0.83)	4.18 (0.75)	2.00 (1.41)	2.73 (1.01)	3.00 (1.41)	3.20
Journalism Support	13	4.00 (0.71)	3.77 (1.01)	1.85 (0.80)	1.38 (0.51)	3.46 (1.20)	2.89
Media Support	10	3.40 (1.07)	3.78 (0.67)	2.00 (0.94)	1.80 (0.92)	3.30 (1.34)	2.86
Taxes / Fines	13	3.08 (1.26)	3.92 (1.32)	2.25 (1.48)	2.83 (1.64)	2.15 (0.99)	2.85
Government Regulation	11	3.64 (1.12)	3.55 (0.82)	1.55 (1.04)	1.82 (0.60)	2.91 (1.14)	2.69
Anti-Trust Action	10	3.60 (1.07)	4.20 (0.63)	1.70 (1.06)	1.40 (0.70)	2.20 (1.40)	2.62
Privacy Legislation	12	3.25 (1.36)	3.80 (0.92)	1.09 (0.30)	1.36 (0.50)	2.75 (1.48)	2.45

Supporting journalists and local media received mediocre scores. Evidence suggests that the decline of local news and especially local newspapers, has contributed to declining civic knowledge, trust, and engagement, [127], which creates conditions for misinformation to spread and increases vulnerability [216]. In the review conducted by Blair et al., they found some initial evidence of the potential effectiveness of training the next generation of journalists. However, they claim that more evidence is needed [46]. Meanwhile, supporting local news organizations could help reverse these trends, but this has not been directly tested, as cost remains an issue when considering the development of quality journalism [37]. Specific interventions could include government subsidies, tax exemptions, or philanthropic support [37].

Surprisingly, investing in local media was also one of the least popular interventions in the Chapter 5 survey, although it was not as unpopular as account suspensions or temporary delays in posting (friction). Considering that this intervention was not perceived as intrusive by users, the unpopularity of this measure stemmed from its perceived unfairness and ineffectiveness. Fairness concerns could be related to the idea that it would support certain media businesses over others. However, investing in local media was also ranked as the second least effective intervention, only above friction. This perception could be addressed by more clearly explaining to users how it addresses the misinformation problem and the potential benefits of investing in quality journalism.

6.4.8 Generative AI

This section presents a comparative analysis of the operationalized generative AI-based interventions (refer to Table 6.2 or Appendix A.8 for detailed definitions).

Objective Characteristics

Table 6.18 shows the objective characteristics of the five operationalized generative AI interventions, listed in the order they were initially presented in Table 6.2.

Table 6.18: Objective characteristics of Generative AI interventions.

Intervention	Implementer	Targeted Phase	Info Changes
Gen AI Chatbots	Platforms, Insts	Belief (verification)	-
Gen AI Content	Platforms, Insts	Creation (content), Belief (verification, prevention)	-
Deepfakes	Platforms, Govts	Creation (content), Spread (amplification)	Removes
AI in Ads	Platforms, Govts	Creation (content), Spread (amplification)	Removes
AI Disclosure	Platforms, Govts	Belief (verification)	Tags

Generative AI interventions are relatively new and share some similarities with other institutional measures or content moderation techniques. The use of chatbots or AI-generated content to produce rebuttals or educational materials can be implemented by either platforms or other institutions, and they are typically aimed at countering belief in misinformation. If external to the platforms, they may not require informational changes to the platforms. Restricting or limiting the use of deepfakes to misrepresent public figures or for advertising purposes specifically aims to combat the creation and spread of misinformation. Meanwhile, requiring clear disclosures or tags indicating that content is AI-generated allows that content to remain on the platforms, and it primarily addresses potential beliefs in misinformation.

Expert Survey Results

Table 6.19 summarizes the expert scores for the generative AI interventions, ranked by highest average score across all five criteria.

Table 6.19: Expert scores for Generative AI interventions. Mean values are shown with their corresponding standard deviations in parentheses.

Intervention	N	Effectiveness	Acceptance	Effort	Cost	Feasibility	Avg
AI Disclosure	14	3.71 (0.99)	4.07 (0.83)	3.31 (1.18)	3.31 (1.25)	3.71 (1.59)	3.62
AI in Ads	18	2.94 (1.00)	3.89 (0.96)	3.00 (1.24)	2.83 (1.34)	3.61 (1.33)	3.26
Deepfakes	11	2.55 (1.29)	4.10 (0.99)	2.40 (1.17)	3.00 (1.25)	4.00 (0.89)	3.21
Gen AI Chatbots	17	3.24 (0.75)	2.71 (0.92)	2.82 (1.01)	2.65 (1.17)	3.39 (0.92)	2.96
Gen AI Content	12	3.00 (0.85)	3.09 (0.83)	2.73 (0.79)	2.64 (0.81)	3.33 (0.65)	2.96

Disclosing the use of AI received the highest score overall. Disclosures and informational interventions, in general, tend to be more popular among the public across various policy domains [118] as well as in the previous chapters of this dissertation. Limiting the use of AI in advertising and prohibiting the use of deepfakes to manipulate the speech or actions of public figures ranked the next highest. However, implementing these interventions may be challenging in practice, as detecting AI can be quite demanding for platforms. Lastly, the use of AI chatbots and AI-generated content ranked the lowest. The usage of AI in such a general way, rather than restricted to specific use cases such as deepfakes or ads, was believed to have low user acceptance. While there is limited research on the effectiveness or acceptance of any generative AI measures, there is some promise regarding the usefulness of chatbots in reducing beliefs in conspiracy theories [78].

6.5 Top-Ranked Interventions

Table 6.20 summarizes the top ten interventions overall, as assessed by misinformation researchers. It also shows the rank of each intervention across the five metrics. This table emphasizes the trade-offs often faced when choosing which interventions to implement. Among the top ten interventions, only three ranked in the top ten for effectiveness, five in acceptance, effort, and cost, and seven in political feasibility. However, most of the top interventions ranked in the top ten for just two or three metrics.

Several interventions not only rank outside the top 10 on multiple metrics but often rank near the bottom in certain categories. For example, user control of one's news feed scores well on user acceptance and feasibility. However, it is ranked 36th out of 40 interventions on effectiveness and is perceived by experts as effortful for platforms to implement (ranking 29th on effort level). Similarly, limiting rampant resharing ranks exceptionally well on effort level, cost, and feasibility, but ranks 35th on user acceptance.

Furthermore, Content Labeling, User-based Measures, and Content Distribution interventions dominate the top 10. Only two interventions from moderation categories (user control and demonetization) rank in the top 10 overall, and these moderation interventions are among the

Table 6.20: Ranking of the top 10 overall interventions by the five metrics.

	Intervention	Effectiveness	Acceptance	Effort	Cost	Feasibility
1.	Government Labels	10	8	2	2	6
2.	Limit Resharing	13	35	1	1	7
3.	Limit Forwarding	17	25	3	3	4
4.	Crowdsourcing	33	6	12	4	2
5.	AI Disclosure	16	7	9	13	11
6.	Friction	25	27	7	6	10
7.	Reporting	11	9	24	20	1
8.	User Control	36	1	29	11	3
9.	Alter Platform Metrics	5	19	20	21	5
10.	Demonetization	3	20	14	25	17

least restrictive or provide users with a high level of agency compared to other moderation interventions. This table highlights the trade-offs often faced when selecting which interventions to implement.

Figure 6.3 presents a heatmap that illustrates the overall average metric scores, averaged by general category and ranked from highest to lowest. Similarly, Table 6.21 shows the aggregated average metric scores by category. Content labeling, user-based measures, and content distribution topped the list among the five criteria. Content labeling and user-based measures scored well on effectiveness, acceptance, and feasibility. Content distribution was the most balanced category with the most similar average metric values and the only category where all five metrics averaged above a 3 on the Likert scale. Media literacy and institutional measures were ranked last, primarily due to their high overall costs and required effort level.

Table 6.21: The average metric values aggregated by intervention category.

	Metric					
Category	Effectiveness	Acceptance	Effort	Cost	Feasibility	Avg
Content Labeling	3.66	3.74	2.81	3.61	3.67	3.50
User-based Measures	3.78	3.86	2.52	3.11	4.03	3.46
Content Distribution	3.49	3.30	3.26	3.48	3.37	3.38
Content Moderation	3.70	3.51	2.66	3.26	3.54	3.33
Generative AI	3.09	3.57	2.85	2.88	3.61	3.20
Account Moderation	3.68	3.10	2.76	3.28	3.10	3.18
Media Literacy / Edu	3.59	4.01	2.05	2.25	3.69	3.12
Institutional Measures	3.55	3.88	1.97	2.05	2.89	2.87

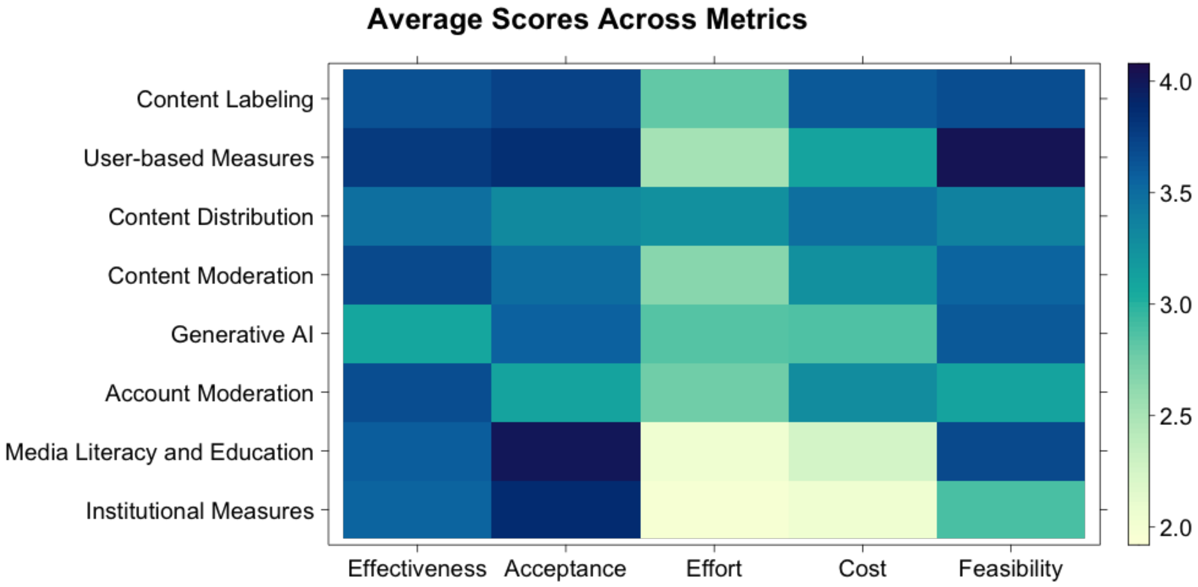


Figure 6.3: A heatmap showing the average score for each metric across general intervention categories.

6.6 Conclusions

6.6.1 Limitations

While this study is among the first expert surveys and the only one so far to analyze multiple evaluative metrics at once, several limitations are noted. First, the sample was limited in both size (n = 39) and geographical reach. The researchers surveyed were based predominantly at U.S.-based academic institutions, rather than industry or non-U.S. institutions, and Carnegie Mellon University researchers in particular represented a significant fraction of the respondents. Second, while the intervention list was expanded to 40, respondents only rated 12 of the interventions to limit the survey length and improve response rates. This limited the sample size for each specific intervention, though all respondents rated the overall effectiveness and acceptance of the general countermeasures categories. Future work should consider a larger sample size or a more diverse set of respondents.

Additionally, the metrics analyzed in this work were equally weighted when comparing and contrasting the different interventions. A limitation of this approach is that some of these metrics could be correlated. An extension of this work could analyze the relationship between the metrics and how that may affect overall ratings. Furthermore, there may be other metrics not included in this work that may be relevant to practitioners or policymakers and could be valuable to investigate in future research. Future research should consider alternative metrics or weighting methods.

6.6.2 Contributions

In this chapter, I characterized and compared the misinformation interventions defined in this dissertation across five evaluative criteria. This analysis incorporated results from previous chapters, existing academic literature, and findings from an expert survey to examine the common features of effective and practical countermeasures and to determine the "best" interventions overall. User-based measures, content labeling, and content distribution scored the highest across a variety of metrics, with content distribution in particular scoring the most consistently across all metrics.

Trade-offs exist when determining which interventions to implement. For example, those high in support, such as media literacy or other educational interventions, often have high costs or implementation effort levels, making them difficult to scale. Additionally, some provably effective interventions, such as content and account moderation techniques, lack user support and acceptance due to their perceived unfairness or intrusiveness. These results highlight the need for platforms to address user concerns, as improving user perceptions could significantly reduce the differences in these evaluative criteria.

Chapter 7

Concluding Remarks

In this thesis, I evaluated the practicality and effectiveness of countermeasures to misinformation and developed a framework to provide analysis-driven recommendations on what to implement and why. I began by comprehensively defining a list of interventions to misinformation in Chapter 1 and conducting a bibliometric analysis of over 400 relevant papers in Chapter 2. Chapters 3 and 4 examined the current state of user-based interventions, user opinions, and strategies to enhance user-based measures and increase participation. Chapter 5 assessed user acceptance of various platform and government interventions and identified the key factors influencing public support. Finally, in Chapter 6, I surveyed misinformation researchers to gather expert opinions on these interventions and integrated that information with the findings from previous chapters to evaluate each category of misinformation interventions.

In this final chapter, I summarize the contributions of this work from a theoretical, methodological, and empirical perspective, and synthesize practical implications for users, platforms, governments, and institutions. Taken together, this work:

- Contributes categorizations relevant to the literature on interventions to counter misinformation
- Develops a theoretical and methodological framework for assessing interventions, and
- Provides meaningful, empirical insights into the effectiveness and user acceptance of countermeasures.

By situating this work within the interdisciplinary and emerging scientific area of social cyber-security [66], I was able to utilize computational social science techniques [111, 217] alongside insights from social network theory [155, 209], and the fields of communications [141], psychology [291], and public policy analysis [118, 248]. I conclude with practical implications and recommendations for future work.

7.1 Contributions

7.1.1 Theoretical Contributions

Chapter 1 explores the various existing definitions and frameworks for categorizing both misinformation and the types of countermeasures. There is a lack of consensus in the literature on these topics, and many challenges are associated with even simply defining misinformation [117]. While explicit falsehoods may be relatively easy to assess or fact-check, misleading or unverified content is more complicated. For example, Facebook initially banned posts that claimed the COVID-19 virus was due to a "lab leak," only to reverse that ban over a year later [132].

This thesis does not address the actual determination or detection of misinformation but instead proposes a general framework for defining types of misinformation more broadly. By articulating a misinformation typology with distinct elements (Agents, Messages, and Audience) and specific features associated with those elements, this work provides a theoretical foundation for the empirical analysis of those misinformation features, building on prior work [305, 317]. The types of agents behind misinformation (such as bots or organizations), the news content of the message (and its potential offline impact), and the targeted audience (such as demographic groups) are all components of social media platform misinformation policies (see Section 1.5). This general typology of misinformation informs the pipeline of misinformation content on social media. Although there is not yet a formal pipeline defined in the literature, previous research indicates that the lifecycle of misinformation content typically involves three phases: the initial network and content creation, the initial spread and subsequent amplification of the content on the platform, and finally, the verification of the content or prevention of false beliefs [72, 212, 305].

Furthermore, the types of countermeasures against social media misinformation were categorized based on standard, objective features of the interventions and informed by prior work in this area [8, 46, 79, 167]. As described in Section 1.5, platforms often define their enforcement policies within these general categories, such as content or account moderation, content labeling, and investing in media literacy. These categories of interventions also typically target specific phases of the misinformation pipeline and, therefore, have different strengths. An important contribution of this work is clearly articulating a pipeline with which one can target interventions.

Finally, the primary theoretical contribution of this thesis is the general framework for assessing and comparing interventions across various metrics, as defined in Chapter 6. The general countermeasures categories defined in Chapter 1 were evaluated based on their objective characteristics and implementation, user perceptions of those characteristics, and five evaluative criteria. While some prior work considers some of these criteria [46, 167, 248], primarily effectiveness and occasionally user acceptance, the combination of multiple criteria in this way is new. This approach opens up many future research directions and provides a framework for other researchers to directly compare interventions.

7.1.2 Empirical Contributions

This dissertation contributes several empirical studies, which mainly use survey analysis to analyze understudied interventions and user acceptance. Chapter 3 provides insights from one of the first large-scale studies of user-based countermeasures to misinformation. Chapter 4 includes a critical assessment of utilizing media literacy training efforts to improve the willingness and ability to counter misinformation, representing one of the first evaluations of its kind. A critical analysis of public opinion about misinformation interventions is conducted in Chapter 5, while expert ratings of these interventions are analyzed in Chapter 6.

Datasets

This thesis additionally contributes several datasets. First, I created a corpus of over 400 papers on interventions to counter misinformation, labeled by both the intervention studied and whether the paper examines user acceptance or the effectiveness of the intervention. This dataset is publicly available on Zotero[1]. It provides other researchers with access to a comprehensive literature review and is easily searchable by keyword or intervention topic label.

Second, this work contributes one of the first large-scale public opinion surveys on user, platform, and government interventions. The data associated with the user-based interventions is published in an open-access article in *Scientific Reports* and is publicly available on the article website [159]. It contains rich demographic data and provides detailed behavioral information on misinformation exposure and responses, broken down by platform and proximity to the misinformation poster. The data associated with the second half of this survey, concerning user perceptions and support levels for various platform and government interventions, will be published elsewhere.

This thesis also includes two other smaller survey datasets. First, I collected rich qualitative data on when people choose to counter misinformation and why from the case study analyzed in Chapter 4. All the social media posts shown in this survey are in Appendix F. Finally, I compiled expert opinions of misinformation researchers regarding interventions in Chapter 6. This survey is among the first to investigate expert opinions on this topic, and it examines one of the most comprehensive lists of interventions (40 operationalized interventions) across five criteria.

Tools

Finally, this work also makes several methodological contributions, primarily the tools used to conduct this research. First, I iteratively designed and tested various ChatGPT prompts to assist in the rapid labeling of hundreds of academic papers by topic. The final prompts can be found in Appendix B. Considering that generative AI is an emerging technology that provides a new type of research assistance, there is no standard method for developing effective prompts. Trial and error, along with informed best practices from other researchers, contributed to the methods used to test these prompts in this thesis, which are described in Chapter 2. I also contribute several Qualtrics survey instruments related to the work in Chapters 3-6, which may help speed the development of future surveys in this research area.

7.2 Summary of Key Findings

This section summarizes the key findings of this dissertation. First, the general findings for user-based measures are summarized in Table 7.1. Chapters 3 and 4 of this thesis explored user-based measures. Chapter 3 conducted a large-scale public opinion survey of American social media users, while Chapter 4 examined detailed qualitative data from a case study conducted on government analysts.

[1] https://www.zotero.org/groups/5961522/misinformation_interventions

#	Finding	Description	Evidence	Ref
1	Hypocrisy	Disparity between actions and beliefs	Ch 3: H1.1 and H2.1	3.3.4
2	Closeness Impact	More likely to counter closer contacts (and oneself) than less close contacts	Ch 3: H1.2 and H3.1 Ch 4: Countering factors	3.3.4 4.5.4
3	Platform Impact	More likely to counter on certain platforms	Ch 3: Platform analysis Ch 4: Countering factors	3.4.2 4.5.4
4	Popularity	Generally popular across most demographic groups	Ch 3: Exploratory results	3.3.4

Table 7.1: Summary of key findings associated with user-based measures.

Table 7.2 summarizes the general findings for platform and government interventions. Chapters 5 and 6 of this thesis explored platform and government-led interventions. Chapter 5 conducted a large-scale public opinion survey of American social media users, asking for their perceptions of various factors related to intervention support, including perceived effectiveness, fairness, and intrusiveness. Chapter 6 explored many of the same questions but posed them to misinformation researchers and expanded the list of interventions and evaluative criteria.

#	Finding	Description	Evidence	Ref
1	User Perceptions	Perceived fairness, effectiveness, and intrusiveness affect support and perceptions	Ch 5: Factors influencing support	5.4.1
2	Preferences	People prefer informational over restrictive interventions	Ch 5: Overall support Ch 6: Top interventions	5.4.2 6.5
3	Support Levels	Partisanship and gender gaps in support	Ch 5: Individual differences	5.4.3
4	Trade-offs	Major trade-offs in five metrics across most interventions	Ch 6: Expert survey results	6.4

Table 7.2: Summary of key findings associated with platform and government interventions.

7.3 General Implications

7.3.1 The Hypocrisy Gap

The first user-based finding reveals a degree of hypocrisy among users. Users often believe they should counter, but they choose not to for various reasons addressed in Chapters 3 and 4. They may believe others will address the problem, feel unprepared or insufficiently knowledgeable to counter, or wish to avoid interpersonal conflict [267, 274]. The qualitative responses from Chapter 4 suggest that addressing these potential concerns could increase the likelihood of users countering themselves or others.

One of the primary implications of this finding is to leverage this gap by implementing public pledges and private reminders to motivate individuals to participate in more pro-social behaviors. Hypocrisy has been shown to lead to behavioral changes in the public health domain, such as increasing intentions to use condoms during the AIDS epidemic [20] and promoting sunscreen use to prevent cancer [271]. It has also been effective in other contexts, such as encouraging water conservation [270], improving respect for speed limits [103], and even reducing racial prejudicial behavior [266].

There is limited research on applying this approach in the social media domain. However, preliminary research from the promotion of a "Pro-Truth Pledge," where users commit to twelve behaviors associated with truthfulness and accuracy, has shown some promising results [284]. The authors note instances where this pledge encouraged signers to retract statements deemed false [284], and a recent experimental study found that truth and sharing discernment improved among signers, as did engagement with verification behaviors [285].

7.3.2 Pipeline Coverage and Intervention Prioritization

The findings related to the platform and government interventions suggest that, like in many other public policy domains, people prefer informational interventions that afford them agency over more intrusive or restrictive ones [87, 123]. On average, content labeling, user-based measures, and content distribution interventions scored the highest among expert raters across the five metrics. When determining which interventions to implement, these top categories unfortunately generally target only the **spread** and **belief** in misinformation rather than the **creation** of networks or content. An ideal enforcement strategy would be to determine and implement the best interventions across the misinformation pipeline.

Among these top three categories, content labeling interventions primarily target content after it has already spread, focusing on the **belief (verification)** phase of the misinformation pipeline. Content distribution interventions generally target the **spread** of misinformation, either addressing the direct sharing component (such as friction and accuracy prompts) or the amplification of the content (by limiting resharing or forwarding). The top-ranked user-based measures included improving user reporting, which also usually targets the **spread (amplification)** phase.

However, two moderation strategies ranked among the top 10 highest-ranked interventions: altering platform metrics to reward accuracy rather than engagement and using demonetization as an account moderation strategy. These interventions target the incentives associated with creating misinformation content and can also affect the potential spread and amplification. Experts in the Chapter 6 survey ranked them among the most effective interventions (5th and 3rd, respectively), although they are perceived to have middling user acceptance and require a decent amount of effort or cost to implement. The literature supports these results, suggesting that reducing incentives to create and share misinformation (or conversely, increasing either monetary or social incentives to share accurate information) can considerably improve the quality of content shared and improve user discernment [113, 150, 290].

7.3.3 Policy Enforcement vs. Perceptions

One of the main findings related to the platform and government interventions is that user perceptions are critical when determining support for misinformation interventions. Fundamentally, platforms handle far more content than can be reasonably audited or moderated, and actual enforcement rates of platform policies are likely unknowable. Even clear-cut illegal content (e.g., CSAM), which platforms often prioritize addressing above other types of violating content, does not achieve 100% enforcement [279].

Several studies have shown that platforms do not uniformly enforce their own rules [133]. For example, Benigni et al. found that account removals were higher for English-speaking ISIS-supporting communities than for other languages [40]. In the period leading up to the 2017 Rohingya genocide, Facebook failed to employ Burmese-speaking moderators and instead relied on translation tools [39, 81]. In another case, Google stopped enforcing its own "warning banner" policy, which was intended to warn users of low-quality search results, just before the 2024 election [211, 246]. Most speech-related policies rely on humans to some degree, which does not easily scale. Algorithmic solutions are often developed only when there is a clear commercial upside [184]. Platforms also tend to prioritize identifying problematic content that has the potential to go viral, while comments or posts in private groups are de-prioritized [133].

Table 7.3 compares the community guidelines enforcement reports of the top social media platforms. While most platforms disclose similar information regarding the volume of content removals categorized by policy violation (such as hate speech, spam, or nudity) and the rates of successful appeals, a few differences emerge. First, only YouTube[2] and TikTok[3] display moderation enforcement by country or language, allowing users to see if policies are equally enforced across markets. TikTok is the only platform that summarizes the removal of various types of fake engagement, the percentage of removals occurring within 24 hours, and the distribution of response times for user-reported content. While Meta's report is comprehensive[4], having several unique pages by policy violation and by platform could hinder information accessibility for the typical user. It is also the only major platform without a specific category for misinformation policy violations. Both Pinterest[5] and TikTok include all the information on one page, though TikTok's report is more interactive and user-friendly.

Public perception of enforcement matters because it can lead to interventions backfiring due to a lack of public support. For example, a large majority of Republicans believe that social media platforms censor certain political views [295]. However, political asymmetries in policy enforcement may be due to differences in misinformation sharing frequency rather than platform bias [204]. Regardless, even if a policy was perfectly enforced and had equal effects across demographic groups, users will lose trust and resist new measures if they perceive it as unfair. In the public opinion survey in Chapter 5, fairness perceptions were more important than both perceived intrusiveness and effectiveness in shaping intervention support.

[2] `https://transparencyreport.google.com/youtube-policy/removals?hl=en` [Accessed 04-30-2025]

[3] `https://www.tiktok.com/transparency/en-us/community-guidelines-enforcement-2024-3` [Accessed 04-30-2025]

[4] `https://transparency.meta.com/reports/community-standards-enforcement/` [Accessed 04-30-2025]

[5] `https://policy.pinterest.com/en/transparency-report` [Accessed 04-30-2025]

Table 7.3: Comparison of enforcement report features across five major social media platforms.

	Platform			
Feature	**YouTube**	**Meta**	**Pinterest**	**TikTok**
Report Frequency	Quarterly	Quarterly	Biannual	Quarterly
Policy Category Breakdown	11 categories	14 categories on Facebook, 12 on Instagram	13 categories	7 categories, several sub-categories per category
Webpage Organization	4 main pages: removals, views, appeals, and flags	Different pages for each category and per platform	One page with all categories	One page with all categories; interactive
Actions Reported	Removals (channels, videos, comments), Appeals (videos)	Removals (content), Appeals (content)	Deactivations and Appeals (pins, boards, accounts)	Removals (comments, LIVE, videos, ads, spam, fake engagement), Restored (videos)
Metrics Reported	View rate	View rate	Reach rate	View rate, Response time distribution,
Detection Method	Automated flagging vs. users vs. orgs/gov	User flagging vs. "found by Meta"	Automated flagging vs. users vs. hybrid	Automated flagging vs. users
Appeals Metrics	Appeals volume and success rates reported overall	Appeals volume and success rates reported by category and platform	Appeals volume and success rates reported by category	Removals and restorations volume; no info on appeals volume
Other Information	Archives since 2018; country/ region breakdown	Archives since 2018; no country/ region breakdown	Archives since 2013; no country/ region breakdown	Archives since 2019; country/ region breakdown

7.3.4 Potential Backfire Conditions

In addition to uneven enforcement, poor implementation or intervention design could affect user perceptions of interventions, potentially leading to backfire conditions. The academic literature has previously investigated potential backfire effects for certain well-studied interventions. Some initial work suggested that fact-checking may have a backfire effect, inducing participants to be more likely to believe the misinformation rather than having no effect on beliefs [215], but that work has not been replicated [234, 311]. Recent work has shown fact-checking to be effective even for people who do not trust fact-checkers [187]. Similarly, serious account moderation techniques, such as deplatforming, were thought to increase the toxicity of accounts on smaller, less moderated platforms [10]. However, those smaller platforms have substantially smaller audiences, resulting in an overall reduction in the reach of these accounts [241].

What is more likely is that certain types of user-based measures or emerging interventions could lead to a backfire effect. For example, media literacy efforts have sometimes been found to be effective [121], but in other cases, depending on implementation, they can lower people's confidence and make them more skeptical of all news, including trustworthy news [298], or even make them more likely to believe the misinformation they were meant to be evaluating [21]. User reporting, if widely advertised to all users, can increase junk reports and be manipulated by organized disinformation groups [99] or weaponized to harass and target specific social media creators [194]. It is still unknown whether Generative AI-based interventions will have a backfire effect. However, in the Chapter 6 survey, experts viewed it as the least effective overall category, with only about 28% of researchers believing it would be effective at reducing misinformation and 46% believing it would be ineffective. This category was the only one with more people disagreeing that it would be effective than agreeing.

7.3.5 Emergence of Generative AI

Generative AI technology has emerged over the last few years as a major player with the power to disrupt existing systems, but also the potential for new interventions. While generative AI-based interventions were included in the expert survey, they emerged after the large-scale public opinion survey conducted for Chapters 3 and 5 was proposed and registered. Nevertheless, the results of this dissertation help inform the discussion on generative AI and misinformation.

First, it remains unclear whether users will be more tolerant of platforms or governments censoring or moderating AI-generated content than human-generated content. However, according to a poll conducted by Pew Research Center in April 2025, about two-thirds of Americans are either very or extremely concerned about AI spreading inaccurate information [192]. This concern holds across partisan lines, with Democrats and Republicans largely aligned on this issue [183]. These results indicate that Americans may be more tolerant of restrictive interventions deployed against AI-generated content than human-generated content.

Additionally, about 60% of Americans are concerned that the government will not go far enough in regulating AI, indicating a desire for stronger oversight [192]. Yet roughly the same percentage of Americans also lack confidence that the government will effectively regulate AI or that companies will develop it responsibly [192]. Americans simultaneously want regulation but are skeptical of any institution's ability to govern AI effectively. The concerns raised in these public opinion polls highlight the need for policymakers to incorporate user feedback while developing transparent, effective, and fair AI moderation and regulatory systems.

This lack of trust and public support also surfaces in the expert survey results in Chapter 6. Only 36% of researchers agreed that these generative AI-based interventions (rebuttals, educational content, chatbots) would be generally acceptable to the American public, making it the category of interventions with the lowest expected user acceptance rates. Because existing opaque algorithmic content moderation and removal strategies are already relatively unpopular (Chapter 5), layering on these additional concerns associated with AI is likely to exacerbate fairness and trust issues.

Despite this general mistrust of AI, some generative AI interventions have already shown promise. For example, chatbots have been used to durably lower beliefs in conspiracy theories [78], and the usage of AI mediation tools has been shown to effectively and fairly help people find common ground [277]. However, in a recent high-profile case, researchers with university IRB approval deployed chatbots on the r/changemyview subreddit. They ran this experiment without disclosing that the accounts were chatbots, prompting outrage from users and the Reddit platform more broadly [161]. It is important to note that similar research was conducted by OpenAI using posts from this subreddit, but without experimenting on non-consenting users [318]. Unfortunately, the lack of regulation and standard ethical practices from companies or researchers may exacerbate the existing mistrust issue with generative AI.

Finally, one of the most transparent ways to address AI-generated misinformation is to visibly and explicitly label AI-generated content. This work is promising, as warning labels more generally have been shown to effectively reduce beliefs in and sharing of misinformation [160, 196, 229]. Early work in this area finds that AI labels are effective at reducing belief and sharing of AI-generated misinformation [102], but they can also make people skeptical of all AI-labeled content, including true stories [15]. Designing an effective AI disclosure intervention involves considering the objectives of the labels, determining what content to label, choosing the appropriate wording for the labels, and considering ways to avoid adverse effects [309].

7.4 Recommendations

This thesis suggests several practical implications and recommendations for social media users, platforms, governments, and other institutions.

7.4.1 For Users

Leverage Social Norms

The findings associated with user-based measures in Table 7.1 indicate that people desire to take more action to combat misinformation than they currently do. Platforms can broadly promote user-based interventions by leveraging social norms, as described in Section 7.3.1, and by promoting “Pro-Truth Pledges” to all users [284], even though some users may be more willing to sign up than others. Users should also educate themselves about misinformation and effective ways to counter it (Chapter 4), and they should strive to hold themselves and those in their communities accountable.

Consider Different User Roles

Not all users engage with social media in the same way, and certain interventions may be more applicable or useful for different user types. Reviewing the literature on social media user roles reveals several common themes, primarily distinguishing between more active user types and those that predominantly lurk [53, 128, 197]. Synthesizing the literature on established user-role frameworks, we can define three actionable roles related to misinformation countering:

- **Spotters** - Users who identify and flag misinformation. This role best suits users who primarily lurk or only sporadically post content [53]. These users comprise the vast majority of social media participants, with previous research finding that only 1% of superusers create most of the content [197]. In Bing He et al.'s recent work on the role of crowds in combatting social media misinformation, these users are defined as "Annotators" who flag or report potentially false or misleading content, playing the critical role of initial detection [128].
- **Verifiers** - Users who evaluate or fact-check content. This role generally falls on users primarily using social media for socializing or debating. According to Brandtzæg's categorization of social media user roles [53], these users frequently engage directly with others by asking questions, challenging assertions, and verifying claims, similar to the "Evaluators" group who assess the validity or effectiveness of corrections or other interventions [128].
- **Creators** - Users who actively create content to counter and correct misinformation. This role suits the most active or advanced users, who generate the majority of content on the platform, including by drafting social corrections or community notes [53, 128, 197].

Platforms can encourage user-based measures by promoting an array of interventions that align with typical user roles, enabling users to utilize the features that most appeal to them. Reporting nudges may be most effective for the Spotters, while quick fact-checks or the encouraging of self or social corrections in post comments could attract the Verifiers. Finally, interactive and user-friendly Community Notes interfaces with reputation incentives might appeal most to the Creators.

7.4.2 For Platforms

Address the Enforcement vs. Perception Gap

Platforms often fail to follow their own policies as defined [133] for various reasons. Because restrictive interventions are sometimes necessary for particularly dangerous or otherwise violating content [81, 279], platforms should be more thoughtful about their policies and the implementation of those policies. The results from the expert survey reveal that moderation techniques, especially those related to account moderation, are typically seen as some of the more effective types of interventions, but they have poor public perception and low user acceptance. Additionally, unlike user-based measures, which are generally supported among most demographic groups, institutional interventions are polarized by gender and partisanship. This political polarization is particularly problematic, as it can affect the feasibility of implementing interventions.

Many of the trade-offs associated with different types of interventions can be addressed by improving user perceptions. For example, trust among the user base can be improved by minimizing errors made by algorithmic detection systems, facilitating a fast and transparent appeals process, and emphasizing fairness throughout [159, 240, 255]. Moderation systems can be improved by investing in scalable hybrid review pipelines that employ automated review systems alongside targeted human review when necessary. Platforms can solicit user feedback on enforcement measures and perceived fairness (e.g., short surveys after a takedown notice or appeals process). Finally, publishing detailed, regular, and easy-to-use transparency reports is critical. These reports should include removal volume by language, country, and policy category; appeal rates and outcomes; and distribution of response and removal times.

Target All Phases of the Misinformation Pipeline

In general, platforms should prioritize informational interventions that are less restrictive when possible (e.g., content labeling). However, they also need to focus on developing effective and practical methods to target the creation phase of the misinformation pipeline. Specifically, changing the incentive structure for creating and sharing misinformation can improve the quality of content on platforms [113, 150, 290]. Experts view altering the fundamental platform metrics or recommendation algorithms that prioritize high engagement to ones that promote a quality information ecosystem as one of the top interventions overall and by effectiveness (Chapter 6). There is strong related evidence in the literature that engagement metrics may make users more vulnerable to misinformation [26], while, conversely, exposure to critical comments by other users on misinformation posts makes users less likely to positively comment or share that misinformation [76]. However, reducing the incentive structure and optimizing for accuracy, or something other than engagement, could undercut the fundamental business model of platforms [37]. These changes may therefore be unlikely to be implemented unless external forces, such as government regulation or changing societal norms, pressure platforms to change.

7.4.3 For Policymakers

Governments play an important role in addressing the misinformation problem, both online and offline. A range of legislation has been passed internationally [77, 104, 214, 248], some of which may not be applicable or politically feasible in the United States [295]. However, policymakers should take note of actions from around the world and consider legislation when appropriate. Promising avenues of government policy in the United States include investment measures, state action, and federal regulation.

Invest in Journalism, Research, and Infrastructure

Federal, state, and local governments should consider supporting local news organizations and journalists through grants, subsidies, or tax exemptions [37, 305]. Additionally, it is important to continue investing in fundamental research to better understand current information issues and potential solutions [130, 305]. Finally, policymakers should work towards improving cybersecurity and resilience of election systems and other infrastructure as a way to neutralize targeted disinformation efforts and influence operations [37, 305].

Consider State Laws as the First Step

When federal action seems unlikely or uncertain, lawmakers should consider prioritizing state-level legislation as an initial step. Several states have enacted various types of privacy legislation [4]. Some have initiated efforts to tax targeted digital advertising [33] or have enhanced media literacy initiatives at the state level [198]. Enacting policy at the state level may allow policymakers to experiment with different approaches, gauge public opinion, and ultimately refine legislation before attempting to implement policy at the national level.

Prioritize Targeted Regulation with Bipartisan Support

Lawmakers and regulators at the federal level should concentrate on targeted regulation or legislation that receives broad bipartisan support and avoids potential censorship concerns. An example of this type of regulation would include the “Honest Ads Act,” which aims to improve the transparency of online political ads [34, 248]. This type of targeted legislation will likely gain broader public support, which is a necessary component for passing legislation [118]. Additionally, policymakers should develop a comprehensive AI regulatory framework and strategy to address the public’s concerns regarding deepfakes and misinformation [130]. There is currently a desire among the American public for regulation in this area [192].

7.4.4 For Institutions

Journalists, researchers, and educators at a variety of civic institutions, including media organizations, think tanks, libraries, universities, and schools, should work collaboratively with industry and government professionals to address the societally pressing issue of misinformation [52, 164, 305].

Advocate for Improved Data Access and Research Funding

Civic organizations are in a unique position to lead collaborations across institutions, advocate for regular citizens, and work towards the development of ethical standards and policy recommendations [52, 164]. Organizations and researchers should advocate for greater transparency from platforms and governments and increased data access to data from consenting users [34, 52]. In recent years, API access for researchers has been increasingly restricted, making it more and more difficult to conduct research in this domain outside of the social media companies [83]. Additionally, the federal government and companies currently fund much of the scientific research in the United States [223]. Increasing research funding from non-profits and other civic organizations would increase flexibility and reduce the reliance on any single source of research funding [52].

Develop Digital Media Literacy Educational Initiatives

Educators, libraries, and schools should work collaboratively to develop age-appropriate curricula focused on news, digital, and AI literacy skills [305]. General critical thinking and research skills, such as the ability to evaluate and assess media sources, are essential. However, educational initiatives should also prioritize incorporating computational and algorithmic literacy into the classroom by helping students understand how algorithms influence and target information.

Governments and civil society organizations can initiate public service campaigns to promote AI media literacy programs and raise awareness of the potential threats associated with deepfakes and generative AI. Additionally, training on open-source intelligence tools for content authentication, such as Google's reverse image search[6], should be expanded [130]. Finally, civic organizations can partner with social media platforms in these efforts by sharing existing digital media literacy educational videos created by platforms, and by collaborating with industry and research partners on identifying the most effective initiatives to implement in the future [34].

7.5 Limitations and Future Research

There are several limitations associated with this research that should be addressed. First, while this work used multiple methods throughout the chapters (bibliometric analysis, public-opinion surveys, qualitative data analysis, and expert ratings), each chapter faced its own restrictions on scope, which limits the generalizability of the results. The literature review in Chapter 2 focused only on English-language papers that were primarily peer-reviewed, potentially overlooking non-English publications or industry work. The empirical studies in the subsequent chapters drew from specific, targeted samples, such as active social media users residing in the U.S. (Chapters 3 and 5), government analysts (Chapter 4), and U.S.-based academic researchers (Chapter 6). These samples do not represent the broader American population. Additionally, self-reported behavior and responses in surveys can be vulnerable to memory recall errors or potential social desirability bias [111]. Finally, to manage survey length, both the public opinion and the expert surveys limited the number of interventions and metrics rated by participants. New and emerging interventions, as well as other potential evaluative metrics, should be assessed in future work.

Second, this dissertation focused primarily on factors and mechanisms previously established as associated with public policy support (perceived effectiveness, fairness, intrusiveness) or behavioral changes (hypocrisy). It did not consider all possible relevant factors, like transparency or problem awareness [118]. It also did not evaluate other potential drivers of opinions or behaviors, such as social identity or analytical reasoning ability [226, 290]. These gaps should be addressed in future work to add to our understanding of which countermeasures and strategies work across a variety of contexts and populations.

Future work associated with user-based measures should investigate how to effectively leverage the hypocrisy gap and better develop interventions for different user types. Inducing hypocrisy through the use of social norms can be more formally tested with additional follow-up studies to the existing work on the effectiveness of the "Pro-Truth Pledge Program" [284, 285]. In partnership with social media teams and other institutions, academics should investigate which types of

[6] `https://images.google.com`

pledges increase initial sign-ups, how durable any behavioral changes are over time, and what factors (such as network factors or individual differences in media literacy or platform trust) have an effect. Studies could be topic-specific (e.g., focused on health or politics) or broad. They could test frequent reminders compared with a single prompt. These studies could help determine the best pledge designs and delivery mechanisms to help close this hypocrisy gap, and how these effects may vary across different user types or social contexts.

When considering institutional interventions, platforms, governments, and institutions should work together to jointly conduct studies on user perceptions, intervention effectiveness, and acceptance across various contexts and user types, while working towards building a comprehensive policy around generative AI. Studies investigating user perceptions of enforcement actions would be best suited to be led by social media teams and other practitioners. This research could explore specific enforcement metrics (e.g., removals by topic/language or successful appeals rates) by running representative user feedback and perception studies to identify gaps in messaging or policy enforcement. Studies external to platforms could examine the effects of intervention intensity or the application of interventions in different contexts, social dynamics, or regional settings to determine if and where backlash and backfire effects may occur.

Finally, experiments determining support for generative AI-based interventions or generative AI content removal should consider several factors. For example, a study could vary the implementer (platform vs. government, as we did in Chapter 5) and content type (human-generated vs. AI-generated) to assess support for AI-generated content removal. Another study could investigate support for AI moderation tools by varying the moderation system used (human-only, AI-only, and hybrid) against measures like trust, perceived fairness, and overall support. Researchers should also investigate the optimal phrasing of AI labels [309] and any potential adverse effects from labeling only a fraction of all AI-generated content.

7.6 Final Comments

This thesis sought to categorize and evaluate different strategies for combatting misinformation on social media platforms. Through this work, I contribute meaningful theoretical frameworks, significant empirical findings on user acceptance of misinformation countermeasures, and a novel approach for comparing and contrasting countermeasures at a high level. The implications of these findings contribute to our broader understanding of this research domain and the greater field of social cyber-security. They also inform actionable recommendations for users, platforms, governments, and institutions to implement. It is important to recognize that there is no one-size-fits-all solution. Certain interventions may be more effective or acceptable for specific types of content, in certain languages or regions, or at particular points in the misinformation pipeline. A diverse arsenal of potential interventions is needed for different situations [37].

Social media empowers average citizens to spread information faster and further than ever before. However, as long as humans have existed, misinformation and rumors have also existed. Addressing the problem of misinformation on social media also has implications for society at large. I hope the work in this thesis contributes to a greater understanding of the way information, especially misinformation, shapes our beliefs and actions, and that it sheds light on potential ways citizens and organizations can combat it.

Appendix A

Countermeasures Categorization

This appendix describes the categorization of specific misinformation interventions developed in Chapter 1, which was used to label the literature in the citation network analysis project described in Chapter 2.

A.1 Content Distribution

Content distribution refers to a broad category of interventions concerning how content is distributed on social media. Specific interventions include:

- **Redirection** - Redirection is 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 [47, 265]. This can be implemented by presenting users with related articles or official information from authoritative sources on the topic, such as the CDC or WHO, when looking for COVID-19 information [79, 315].
- **Accuracy Prompts** - Sometimes referred to as "nudging", accuracy prompts are a type of content distribution designed to encourage individuals to consider accuracy before posting or sharing content [231]. These prompts can be implemented in various ways, such as by asking users to evaluate the accuracy of one or more headlines or to reflect on the importance of sharing accurate news before continuing on the platform [101].
- **Friction** - Friction encourages individuals to pause and reflect before engaging with content [29, 151]. This can be operationalized by temporarily preventing users from resharing content they have not opened, prompting them to consider reading the article through interstitials or pop-up windows [236].
- **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 [160]. Examples include reducing the size or visibility of a post [109, 160, 262] or altering platform architecture to guide users to move posts to specific rather than general groups [156].
- **Advertising Policy** - Advertising policy refers to how platforms (or governments) regulate, correct, or display advertisements to the public [34, 79, 131]. Examples include banning

political ads, requiring ads to undergo a fact-checking process before posting, prohibiting certain fake news websites from advertising, or labeling ads as “paid for” [9, 70, 79].

- **Other Content Distribution** - Other types of content distribution involve limiting users’ forwarding capabilities, which caps the number of recipients to whom a given message can be forwarded [235], and limiting resharing, which restricts rampant resharing by removing share buttons on posts after several levels of sharing [6].

A.2 Content Moderation

Content moderation refers to a broad range of interventions regarding the way content is displayed or not displayed on social media. This encompasses fact-checking, narrative counter-speech such as debunking, and the use of algorithms to assist in moderation and misinformation detection [143, 257]. Specific interventions include:

- **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 [50, 300].
- **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” [68, 182].
- **Algorithmic Content Moderation** - This refers to automated content moderation, including automated fact-checking, labeling, or removing posts containing misinformation or other violating content. It can also involve downranking or de-emphasizing low-quality content, as well as upranking or promoting high-quality or authoritative news sources [34, 48, 109, 116]. Additionally, it can be employed as a virality circuit breaker, which automatically flags certain fast-spreading and unverified content, temporarily halting algorithmic amplification until the information is verified [6].
- **Misinformation Detection** - The algorithmic detection of misinformation, typically for content moderation purposes [153, 185].
- **Other Content Moderation** - 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 [142].

A.3 Account Moderation

Account moderation involves moderating user accounts by implementing actions such as account suspensions, removals, shadowbanning users, or demonetizing accounts [79, 308]. Specific interventions include:

- **Account Removal** - 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 [241, 280].

- **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 [144, 308].
- **Demonetization and Other Account Moderation** - Another form of account moderation is the demonetization of user accounts [190]. This process refers to removing or restricting monetization features for a user account that is found to violate a platform's policies repeatedly.

A.4 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 [202, 315]. Specific interventions include:

- **Crowdsourcing** - Crowdsourcing generally involves utilizing ordinary individuals to verify information and label content instead of relying on journalists or expert fact-checkers [12]. The most recognized operational version of a crowdsourcing intervention is X's Community Notes program [13].
- **Warning Labels** - Warning labels refer to general warnings about misinformation and can address the source, content, or context [222]. One way to implement this type of intervention is by using click-through warning labels or interstitials, which both warn users and encourage them to reflect before viewing the content [262].
- **Source Credibility Labels** - This type of intervention involves disclosing or labeling the credibility of a post's source. This intervention can be implemented by labeling the reliability of a news source[105] or by labeling the accounts of government officials or state-run media sources [208].
- **Context Labels** - Context labels are labels that specifically add context or additional information to a post, such as Community Notes programs [13].
- **Other Content Labeling** - A related type of content labeling involves specifically notifying users when they have posted or interacted with content verified to contain misinformation or originating from a state-run media source through the use of misinformation disclosure [79].

A.5 User-based Measures

User-based measures are interventions that focus on individuals' responses to encountering misinformation [274]. Specific interventions include:

- **Reporting** - Users can report other users or their posts [214, 321].
- **Blocking** - Users can block other users or specific topics [274].

- **Social Corrections** - Users can fact-check or debunk other users directly [28, 48]. This intervention may involve publicly commenting on a post or privately messaging the misinformation poster.
- **Social Norms** - The user community has tools to influence user behavior and promote social and self-corrections [98, 110]. This could be operationalized by adjusting platform metrics to reward accuracy instead of engagement, aiming to discourage the sharing of misinformation [26].
- **Retractions** - Users or organizations can retract misinformation they have posted and try to mitigate the impacts on individuals who have already encountered the misinformation [18, 95].
- **Other User-based Measures** - Users can access tools for self-moderation, behavior change, and deactivation of their social media accounts in response to nudges [292].

A.6 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 [121, 141]. Specific interventions include:

- **Fake News Games** - Games designed to help players detect misinformation and improve their critical thinking skills [186, 200, 250].
- **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 [180].
- **Other Media Literacy Efforts** - Other types of educational initiatives and relevant research studies, such as providing individuals with tips on recognizing fake news [121], evaluating people's information, digital, or news literacy [146], or developing strategies to improve media literacy messaging [288].

A.7 Institutional Measures

Institutional measures are those implemented by civic society, governments, the media, or other organizations [52]. Specific interventions include:

- **Media Support** - Investing in local news or promoting trustworthy local news on social media platforms [281]. Supporting and training the next generation of journalists to engage in high-quality, independent reporting [32, 52].
- **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 [22, 34, 44].
- **Government Regulation** - This category includes any laws, rules, or regulations at local, state, or federal levels [214, 248, 314]. Regulation may involve holding companies

accountable for the content shared on their platforms, such as by modifying Section 230 or regulating them like utility or media companies. Legislation could involve developing comprehensive privacy laws, similar to Europe's GDPR, or taking anti-trust action by breaking up monopolistic technology companies [248]. Furthermore, companies could be taxed or fined for using personal user data, and micro-targeted advertising could be banned or restricted [237].

- **Other Institutional Measures** - Additional institutional measures may involve researching and developing tools to support civil society with these issues, as well as enhancing collaboration among various types of institutions [52, 305].

A.8 Generative AI and Other

This category includes any interventions that do not fit into the previous categories or are newly introduced. Specific interventions include:

- **Generative AI** - Using generative AI to combat or detect misinformation. Examples include employing AI chatbots to reduce belief in conspiracy theories or misinformation [78] and using AI-generated content to create rebuttals against misinformation or to develop educational initiatives. Policy responses may involve prohibiting or labeling the use of AI or manipulated content to produce deepfakes, banning the use of AI in advertising, or requiring clear disclosure on any ads that incorporate AI-generated images, videos, or audio [130].
- **Combining Interventions** - Studies that explicitly compare the effects of using multiple interventions simultaneously with using one intervention [30].
- **Other** - Any other intervention not previously described.

Appendix B

Chapter 2: ChatGPT Prompts

This appendix includes the final ChatGPT prompts that were created to assist with the labeling of papers in Chapter 2.

B.1 Effectiveness Labeling Prompt

Effectiveness Labeling - Task Overview:

You are an academic researcher tasked with labeling research papers (PDFs) focused on interventions aimed at countering misinformation. Your goal is to determine whether each paper studies the effectiveness of intervention(s) or action(s) in reducing misinformation.

Instructions for Labeling:

Label as "Yes" (Studies Effectiveness):

Mark papers as "Yes" if they empirically measure or analyze, either directly or indirectly, the effectiveness of an intervention in reducing the creation, belief, sharing, or spread of misinformation. This includes studies that assess effectiveness through:

- Direct measurement, such as experimental data or controlled trials, or
- Indirect analysis, such as observational studies or secondary analyses that provide insights into intervention impact
- Papers that find an intervention is not effective should still be labeled "Yes" for studying effectiveness

Label as "No" (Does Not Study Effectiveness):

Mark papers as "No" if they do not empirically evaluate the effectiveness of an intervention, directly or indirectly.

- This includes theory-driven papers focused on conceptual frameworks or hypotheses without testing outcomes, as well as discussion-based papers that explore ideas, potential strategies, or frameworks without empirical testing.
- Literature reviews that summarize existing research without conducting empirical analysis related to effectiveness also fall under this category.

B.2 Acceptance Labeling Prompt

Acceptance Labeling - Task Overview:

You are an academic researcher tasked with labeling research papers (PDFs) focused on interventions aimed at countering misinformation. Your goal is to determine whether each paper studies the public's or users' acceptance or support of intervention(s) or action(s) to counter misinformation.

Instructions for Labeling:

Label as "Yes" (Studies Acceptance):

Mark papers as "Yes" if they explicitly explore users' or the public's opinions, perceptions, or acceptance of an intervention. Papers that qualify for this label often:

- Include surveys, interviews, or other methods of feedback collection on user preferences.
- Explore potential improvements to the intervention based on user feedback.

Label as "No" (Does Not Study Acceptance):

Mark papers as "No" if they do not examine user acceptance or support. Papers that should be labeled "No" often:

- Focus exclusively on the effectiveness, impact, technical performance, or societal impact of the intervention without assessing user feedback
- Some papers may include surveys or interviews but still do not qualify if these methods are used solely to analyze intervention performance, not user acceptance.

B.3 Interventions Labeling Prompt

Interventions Labeling - Task Overview:

You are an academic researcher tasked with labeling research papers (PDFs) focused on interventions aimed at countering misinformation. Your goal is to determine the primary intervention(s) or action(s) studied in each paper.

Instructions for Labeling:

- Assign only the most relevant label(s) that best capture the central intervention(s) investigated in each paper.
- Do not label interventions that are only briefly mentioned. Only label interventions that are a key focus of the study.
- If a paper discusses multiple interventions, prioritize the most rigorously studied or most emphasized intervention(s).
- Use only the labels provided below and format them in a comma-separated list.
- If no label fits, use "other" instead of forcing an inappropriate label.

Label Definitions: Interventions are broken up by category for clarity.

- Content Distribution interventions relate to the methods by which content is distributed on social media.
  - redirection: directing users to alternative or official content, or showing related content.
  - accuracy prompts: reminding users to assess content accuracy before sharing.
  - friction: adding delays, prompts, or other mechanisms to slow down engagement and encourage thoughtful interaction.
  - platform alterations: structural changes affecting how content is distributed or displayed to users.
  - advertising policy: rules governing advertisements and their effects.
  - content distribution: unspecified content distribution intervention that is not described by one of the other labels.
- Content Moderation interventions determine what content is displayed or suppressed.
  - fact-checking: verifying information accuracy, often by expert fact-checkers or journalists.
  - debunking: providing context or narrative coherence to correct or refute misinformation.
  - misinformation detection: using algorithmic techniques to identify misinformation.
  - algorithmic content moderation: the use of automated moderation techniques.
  - content moderation: unspecified content moderation intervention that is not described by one of the other labels.
- Account Moderation interventions are policies and actions to moderate user accounts.
  - account removal: banning users or coordinated deplatforming.
  - shadow banning: restricting account or post visibility without an official ban.
  - demonetization: restricting or removing monetization options from a user or content.
  - account moderation: unspecified account moderation intervention that is not described by one of the other labels.
- Content Labeling interventions are labels that disclose or provide context for misinformation.
  - crowdsourcing: user-generated contributions that help label or assess misinformation.
  - warning labels: explicit labels warning users about misinformation.
  - source credibility labels: labels that indicate an information source's reliability.
  - context labels: adding background or contextual information to a post.
  - content labeling: unspecified content labeling intervention that is not described by one of the other labels.
- User-based interventions are actions users can take to respond to misinformation.
  - reporting: enabling users to flag misinformation for review.

  - blocking: user action that prevents engagement with specific accounts.
  - social norms: encouraging platform-specific social or community-based norms that discourage misinformation.
  - social corrections: direct corrections provided by other users in response to misinformation.
  - retractions: users voluntarily retracting or correcting their own misinformation.
  - user-based measures: unspecified user-based intervention that is not described by one of the other labels.
- Media Literacy interventions are educational efforts for critical media engagement.
  - fake news games: interactive games designed to train users to recognize misinformation
  - inoculation: pre-bunking techniques that prepare users to recognize and resist misinformation.
  - media literacy: unspecified media literacy intervention that is not described by one of the other labels.
- Institutional interventions are interventions by organizations or authorities.
  - media support: supporting or promoting reliable, high-quality journalism and local news sources.
  - data sharing: enabling ethical data access for misinformation research.
  - government regulation: laws or policies designed to regulate misinformation or social media platforms.
  - institutional measures: unspecified institutional intervention that is not described by one of the other labels.
- Other interventions are miscellaneous or multiple combined interventions.
  - generative AI: AI usage to counter misinformation.
  - combining interventions: analyzing the use of multiple intervention strategies at once instead of one at a time.
  - other: interventions that do not fit any of the above categories.

Appendix C

Chapter 3: Pre-Registered Hypotheses and Analyses

The design table summarizes the research questions, hypotheses, power calculations, planned analyses, and possible interpretations for the closeness analysis in Chapter 3.

Question	**Details**
RQ1: How do people respond and think others should respond when they see misinformation?	**H1.1:** People believe individuals should expend more effort to respond to misinformation online than they actually do. **Sampling Plan:** The necessary sample size for a one-sided paired Bayesian test to achieve 95% power to detect a medium effect size of 0.5 at a Bayes threshold of 10 is estimated to be 81. The estimated number of participants who have seen misinformation per closeness level is approximately 192. **Analysis Plan:** • Calculate Measures 1a and 1b at each of the three closeness levels. • Run three one-sided Bayesian paired hypothesis tests (one at each closeness level). • Calculate the 95% highest density interval (HDI) for the effect size. • Visualize results with bar charts or similar visualizations comparing possible actions. **Interpretation:** The resulting Bayes factor will inform the probability of the alternate or null hypothesis being true. For example, a Bayes factor of 10 implies the data is 10 times more likely under H1 than H0.

Question	Details
RQ1: How do people respond and think others should respond when they see misinformation?	**H1.2:** 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 **Sampling Plan:** The necessary sample size for a one-sided paired Bayesian test to achieve 95% power to detect a medium effect size of 0.5 at a Bayes threshold of 10 is estimated to be 81. The estimated sample size is approximately 115. **Analysis Plan:** • Calculate Measures 1a at each of the three closeness levels. • Run a one-sided Bayesian paired hypothesis test comparing the effort level (Measure 1a) for close contacts and somewhat close contacts. Similarly, run a one-sided Bayesian paired hypothesis test comparing the effort level (Measure 1a) for somewhat close contacts and not close contacts. • For both tests, calculate the 95% highest density interval (HDI) for the effect size. Visualize the results. **Interpretation:** The resulting Bayes factor will inform the probability of the alternate or null hypothesis being true. Visualizations will help the reader interpret the results.
RQ1: How do people respond and think others should respond when they see misinformation?	**H1.3:** People ***believe*** others should 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. **Sampling Plan:** The necessary sample size for a one-sided paired Bayesian test to achieve 95% power to detect a medium effect size of 0.5 at a Bayes threshold of 10 is estimated to be 81. All participants will be included in this test, indicating a sample size of approx. n = 1000. **Analysis Plan:** • Calculate Measures 1b at each of the three closeness levels. • Run a one-sided Bayesian paired hypothesis test comparing the effort level (Measure 1b) for close contacts and somewhat close contacts. Similarly, run a one-sided Bayesian paired hypothesis test comparing the effort level (Measure 1b) for somewhat close contacts and not close contacts. • For both tests, calculate the 95% highest density interval (HDI) for the effect size. Visualize the results. **Interpretation:** The resulting Bayes factor will inform the probability of the alternate or null hypothesis being true. Visualizations will help the reader interpret the results.

Question	Details
RQ2: How do people behave after realizing they have posted misinformation?	**H2.1:** People believe others should expend more effort to respond to misinformation online after realizing they posted misinformation than they actually do. **Sampling Plan:** The necessary sample size for a one-sided paired Bayesian test to achieve 95% power to detect a medium effect size of 0.5 at a Bayes threshold of 10 is estimated to be 81. The estimated number of participants who have posted misinformation accidentally is n = 320. **Analysis Plan:** • Calculate Measures 2a and 2b to determine response effort. • Run a one-sided Bayesian paired test with these two measures. • Calculate the 95% HDI for the effect size and visualize the results. **Interpretation:** The resulting Bayes factor will inform the probability of the alternate or null hypothesis being true. The visualizations will help readers understand the effort differences.
RQ3: How do people's responses and beliefs about how others should respond after seeing misinformation differ from their responses and beliefs when they realize they have posted misinformation?	**H3.1:** People respond with a different level of effort when the sender of misinformation is someone else compared to themselves. **Sampling Plan:** The necessary sample size for a two-sided paired Bayesian test to achieve 95% power to detect a medium effect size of 0.5 at a Bayes threshold of 10 is estimated to be 92. The estimated number of participants is 150. **Analysis Plan:** • Calculate Measures 1a (for each closeness level) and 2a. • Run a two-sided paired Bayesian test comparing these two measures. • The null hypothesis is these two measures are equal. The null interval is -0.2 to 0.2 (effect size is effectively 0). The alternate hypothesis is that the absolute value of the effect size is at least 0.2. • Calculate the 95% HDI and visualize the results. **Interpretation:** The resulting Bayes factor will inform the probability of the alternate or null hypothesis being true. If the 95% highest density interval falls entirely within the "region of practical equivalence" (ROPE) range, of -0.2 to 0.2, then we can accept the null. If it falls entirely outside the range, then we can accept the alternate. Otherwise, using this method we will state inconclusive results. The visualizations will help readers understand the effort differences.

Question	Details
RQ3: How do people’s responses and beliefs about how others should respond after seeing misinformation differ from their responses and beliefs when they realize they have posted misinformation?	**H3.2:** People want others to respond with a different level of effort when the sender of misinformation is someone else compared to themselves. **Sampling Plan:** The necessary sample size for a two-sided paired Bayesian test to achieve 95% power to detect a medium effect size of 0.5 at a Bayes threshold of 10 is estimated to be 92. All participants will be included in this test, indicating a sample size of approx. n = 1000. **Analysis Plan:** • Calculate Measures 1b (for each closeness level) and 2b. • Run a two-sided paired Bayesian test comparing these two measures. • The null hypothesis is these two measures are equal. The null interval is -0.2 to 0.2 (effect size is effectively 0). The alternate hypothesis is that the absolute value of the effect size is at least 0.2. • Calculate the 95% HDI and visualize the results. **Interpretation:** The resulting Bayes factor will inform the probability of the alternate or null hypothesis being true. If the 95% highest density interval falls entirely within the “region of practical equivalence” (ROPE) range, of -0.2 to 0.2, then we can accept the null. If it falls entirely outside the range, then we can accept the alternate. Otherwise, using this method we will state inconclusive results. The visualizations will help readers understand the effort differences.

Appendix D

Chapter 3: Supplemental Categorical Analysis

Categorical analysis was conducted as a robustness check to better understand the association between the three categorical variables in the model: **effort level** when countering, **response type** (reported behavior vs. opinion), and **closeness level**. Only the beliefs of participants who saw misinformation at each closeness level are included to ensure this analysis is analogous to those of our pre-registered hypothesis tests. Considering the different actions available for countering oneself (Table 3.2) versus others (Table 3.1), we first only consider responses from others. Table D.1 shows the three-way contingency table, which underpins Figure 3.2 in Chapter 3. The generalized McNemar's chi-squared tests for categorical paired data were run to assess the relationship between effort level and response type at each closeness level, and all yielded statistically significant results. These findings are consistent with our results from H1.1.

		Response Type		**McNemar's**
Closeness Level	**Effort Level**	Behavior	Opinion	**Chi-sq. Test**
Close contacts (n = 148)	No Effort	52	20	$\chi^2 = 34.9^{***}$
	Low Effort	19	13	$p = 1\text{e-}7$
	High Effort	77	115	
Somewhat close contacts (n = 370)	No Effort	179	76	$\chi^2 = 126.3^{****}$
	Low Effort	76	41	$p < 2\text{e-}16$
	High Effort	115	253	
Not close contacts (n = 880)	No Effort	462	248	$\chi^2 = 201.5^{***}$
	Low Effort	217	251	$p < 2\text{e-}16$
	High Effort	201	381	

Table D.1: Contingency table of effort level, response type, and closeness level. Each cell contains the frequency of occurrences for each combination of the categorical variables. The chi-sq test statistic is shown: $p < 0.05^{*}$, $p < 0.01^{**}$, and $p < 0.001^{***}$.

Since the contingency table includes some repeated responses from the same participants, we analyzed the data using generalized estimating equations (GEE) with the “geepack” R package [137] to estimate the average population-level effects of the categorical variables. Table D.2 shows the two best model results from the analysis based on QIC values, which are modified AIC values applicable to GEE models [220]. We compared three models to predict the count frequencies in the contingency table: the model with main effects only, the model with main effects and two-way interaction terms, and the saturated model with the three-way interaction term. According to the QIC values, the saturated model with the three-way interaction term was the best, even though the three-way interaction term was not statistically significant. Several two-way interaction terms are significant. There is a higher frequency of “high effort” responses when contacts are closer or when people are asked about their opinions instead of their actual behavior, echoing our results from RQ1.

We conducted a similar analysis of the responses against oneself. Table D.3 shows the two-way contingency table and the generalized McNemar’s chi-squared test result, which was statistically significant. This finding is consistent with our result from H2.1. Table D.4 shows the model to predict the frequency of counts in the “self” contingency table. Again, effort level interacts with response type, with higher effort actions seeing higher counts when people are asked about their opinion rather than their actual behavior, echoing our results from H2.1. We also see higher counts for low and high-effort actions compared to no-effort actions. The estimates in this model for low and high effort counts are also higher than the comparable estimates in Table D.2, indicating that individuals may be exerting more effort to correct themselves than others.

Table D.2: The generalized estimating equation (GEE) model predicting the average frequency counts as a function of effort level (reference level: no effort), response type (reference level: behavior), and closeness (reference level: close contacts) when considering responses against others.

	Dependent variable: Frequency counts			
	Model with Three-Way Terms		Model with Two-Way Terms	
	Estimate	Std. Err.	Estimate	Std. Err.
Low Effort	-1.007^{***}	(0.264)	-1.038^{***}	(0.213)
High Effort	0.393^{*}	(0.173)	0.380^{**}	(0.143)
Opinion	-0.956^{***}	(0.259)	-0.994^{***}	(0.132)
Not Close	2.184^{***}	(0.139)	2.179^{***}	(0.123)
Somewhat Close	1.236^{***}	(0.151)	1.210^{***}	(0.134)
Low Effort : Opinion	0.576	(0.441)	0.664^{***}	(0.091)
High Effort : Opinion	1.357^{***}	(0.294)	1.398^{***}	(0.082)
Low Effort : Not Close	0.251	(0.273)	0.339	(0.216)
High Effort : Not Close	-1.225^{***}	(0.187)	-1.299^{***}	(0.149)
Low Effort : Somewhat	0.150	(0.294)	0.003	(0.237)
High Effort : Somewhat	-0.835^{***}	(0.205)	-0.676^{***}	(0.160)
Opinion : Not Close	0.333	(0.267)	0.358^{**}	(0.128)
Opinion : Somewhat	0.099	(0.289)	0.188	(0.137)
Low Effort : Opinion : Not Close	0.192	(0.453)		
High Effort : Opinion : Not Close	-0.095	(0.310)		
Low Effort : Opinion : Somewhat	-0.337	(0.496)		
High Effort : Opinion : Somewhat	0.288	(0.337)		
Constant	-2.882^{***}	(0.135)	-2.871^{***}	(0.120)

Note: *p<0.05; **p<0.01; ***p<0.001

		Response Type		**McNemar's**
	Effort Level	Behavior	Opinion	**Chi-sq. Test**
Self (n = 256)	No Effort	13	5	$\chi^2 = 63.2^{***}$
	Low Effort	114	52	$p =$ 1e-13
	High Effort	129	199	

Table D.3: Contingency table of effort level and response type for actions against oneself. Each cell contains the frequency of occurrences for each combination of the categorical variables. The chi-sq test statistic is shown: $p < 0.05^{*}$, $p < 0.01^{**}$, and $p < 0.001^{***}$.

Table D.4: The generalized estimating equation (GEE) model predicting the average frequency counts as a function of effort level (reference level: no effort) and response type (reference level: behavior) when considering responses to oneself.

	Dependent variable: Frequency counts	
	Model with Two-Way Terms	
	Estimate	Std. Err.
Low Effort	2.171^{***}	(0.279)
High Effort	2.295^{*}	(0.277)
Opinion	-0.956	(0.519)
Low Effort : Opinion	0.171	(0.538)
High Effort : Opinion	1.389^{**}	(0.524)
Constant	-2.980^{***}	(0.270)
Note:	*p<0.05; **p<0.01; ***p<0.001	

Appendix E

Chapter 3: Demographic Analysis

Table E.1 shows the percentage of each demographic category that believes one should respond with a maximum of no, low, or high effort when encountering misinformation posted by others or oneself. Any category with a sample size of fewer than 50 participants was combined with the next closest demographic category, except the "Other" category for gender. The chi-square test of independence was conducted for each demographic category and each closeness level. This test is used to determine whether the beliefs of the demographic groups are independent of one another.

Table E.1: Demographics summary table.

Category	**Sample Size (n)**	**% Effort (No / Low / High)**				**Chi-sq Test of Independence Results**
		Close contacts	**Somewhat close**	**Not close**	**Self**	
Overall	n = 1010	14.5 / 5.2 / 80.3	18.2 / 8.1 / 73.7	29.2 / 27.0 / 42.8	2.5 / 19.2 / 78.3	$p < 0.05^{*}$, $p < 0.01^{**}$, $p < 0.001^{***}$
Gender	Female (465)	13.3 / 4.3 / 82.4	17.0 / 8.82 / 74.2	28.0 / 26.9 / 45.2	1.7 / 18.3 / 80	**Test statistic:** Close: 3.02
	Male (520)	16.0 / 6.0 / 78.1	20.0 / 7.5 / 72.5	31.2 / 28.5 / 40.4	3.3 / 20 / 76.7	Somewhat: 1.82
	Other (25) *(excl. from stat. tests)*	4.0 / 8.0 / 88.0	4.0 / 8.0 / 88.0	12.0 / 40.0 / 48.0	0 / 20 / 80	Not Close: 2.38, Self: 3.03

Category	Sample Size (n)	% Effort (No / Low / High)				Chi-sq Test of Independence Results
		Close contacts	Somewhat close	Not close	Self	
Age	18-34 (253)	21.3 / 7.5 / 71.1	24.1 / 9.5 / 66.4	32.4 / 29.6 / 37.9	4.7 / 23.7 / 71.5	**Test statistic:**
	35-44 (338)	13.3 / 5.9 / 80.8	18.0 / 10.9 / 71.0	24.0 / 31.4 / 44.7	2.7 / 20.7 / 76.6	Close: 25.5**
	45-54 (186)	11.8 / 3.8 / 84.4	15.1 / 4.8 / 80.1	30.1 / 25.8 / 44.1	1.1 / 17.7 / 81.2	Somewhat: 26.4***
	55-64 (148)	14.2 / 2.7 / 83.1	18.9 / 5.4 / 75.7	33.1 / 23.6 / 43.2	1.0 / 12.2 / 87.2	Not Close: 10.9
	65+ (85)	4.7 / 3.5 / 91.8	7.1 / 4.7 / 88.2	31.8 / 22.4 / 45.9	1.2 / 15.3 / 83.5	Self: 20.7**
Race	White/ Caucasian (799)	14.5 / 5.6 / 79.8	18.5 / 7.9 / 73.6	30.3 / 26.2 / 43.6	2.4 / 19.9 / 77.7	**Test statistic:** Close: 8.88
	Black or African American (85)	9.4 / 4.7 / 85.9	15.3 / 10.6 / 74.1	21.2 / 31.8 / 47.1	2.4 / 17.6 / 80.0	Somewhat: 2.67
	Asian (65)	10.8 / 3.1 / 86.2	15.4 / 6.2 / 78.5	21.5 / 41.5 / 36.9	3.1 / 18.5 / 78.5	Not Close: 12.64*
	Multiracial/ Other (61)	24.6 / 3.3 / 72.1	21.3 / 9.8 / 68.9	34.4 / 32.8 / 32.8	3.3 / 13.1 / 83.6	Self: 2.08
Education	High school or less (116)	17.2 / 1.7 / 81.0	22.4 / 4.3 / 73.3	32.8 / 25.9 / 41.4	3.5 / 21.6 / 75.0	**Test statistic:**
	Some college (164)	8.5 / 3.1 / 88.4	15.9 / 6.1 / 78.0	20.7 / 22.6 / 56.7	1.2 / 15.2 / 83.5	Close: 17.2*
	Associate’s degree (126)	11.9 / 4.0 / 84.1	15.9 / 5.6 / 78.6	27.8 / 27.0 / 45.2	1.6 / 23.0 / 75.4	Somewhat: 9.9
	Bachelor’s degree (438)	16.2 / 6.2 / 77.6	18.7 / 9.6 / 71.7	31.1 / 30.6 / 38.4	3.0 / 19.9 / 77.2	Not Close: 18.5*
	Master’s degree or higher (166)	15.7 / 8.4 / 75.9	18.1 / 10.8 / 71.1	31.3 / 28.9 / 39.8	2.4 / 16.9 / 80.7	Self: 6.6
Income	Less than $20,000 (87)	10.3 / 4.6 / 85.1	11.5 / 5.8 / 82.8	21.8 / 28.7 / 49.4	0 / 18.4 / 81.6	
	$20,000 - $39,999 (172)	11.0 / 4.7 / 84.3	12.8 / 6.4 / 80.8	23.3 / 31.4 / 45.3	2.9 / 21.5 / 75.6	**Test statistic:**
	$40,000 - $59,999 (220)	15.9 / 3.6 / 80.5	19.1 / 5.5 / 75.5	28.6 / 22.7 / 48.6	3.6 / 18.6 / 77.7	Close: 10.2
	$60,000 - $79,999 (181)	13.8 / 5.5 / 80.7	17.7 / 9.9 / 72.4	26.5 / 28.7 / 44.8	2.2 / 11.6 / 86.2	Somewhat: 25.1*
	$80,000 - $99,999 (140)	14.3 / 6.4 / 79.3	23.6 / 11.4 / 65.0	37.1 / 27.1 / 35.7	2.9 / 18.6 / 78.6	Not Close: 22.4*
	$100,000 - $149,999 (131)	16.0 / 7.6 / 76.3	16.8 / 11.5 / 71.8	33.6 / 33.6 / 32.8	3.1 / 22.1 / 74.8	Self: 20.2
	Over $150,000 (79)	21.5 / 5.1 / 73.4	29.1 / 6.3 / 64.6	36.7 / 25.3 / 38.0	0 / 30.4 / 69.6	

Category	Sample Size (n)	% Effort (No / Low / High)				Chi-sq Test of Independence Results
		Close contacts	Somewhat close	Not close	Self	
Party	Strong Republican (130)	20.8 / 6.9 / 72.3	24.6 / 5.4 / 70.0	33.8 / 23.8 / 42.3	5.4 / 23.1 / 71.5	**Test statistic:**
	Weak Republican (106)	14.2 / 1.9 / 84.0	19.8 / 4.7 / 75.5	35.8 / 26.4 / 37.7	1.0 / 25.5 / 73.6	Close: 25.4**
	Independent/ Other (303)	15.8 / 2.0 / 82.2	20.1 / 7.3 / 72.6	33.7 / 24.1 / 42.2	2.0 / 17.2 / 80.9	Somewhat: 14.1
	Weak Democrat (203)	12.8 / 5.4 / 81.8	15.8 / 8.4 / 75.9	29.6 / 30.5 / 39.9	2.5 / 18.7 / 78.8	Not Close: 22.3**
	Strong Democrat (268)	11.2 / 9.3 / 79.5	14.2 / 11.6 / 74.3	19.0 / 33.2 / 47.8	2.2 / 17.5 / 80.2	Self: 11.5

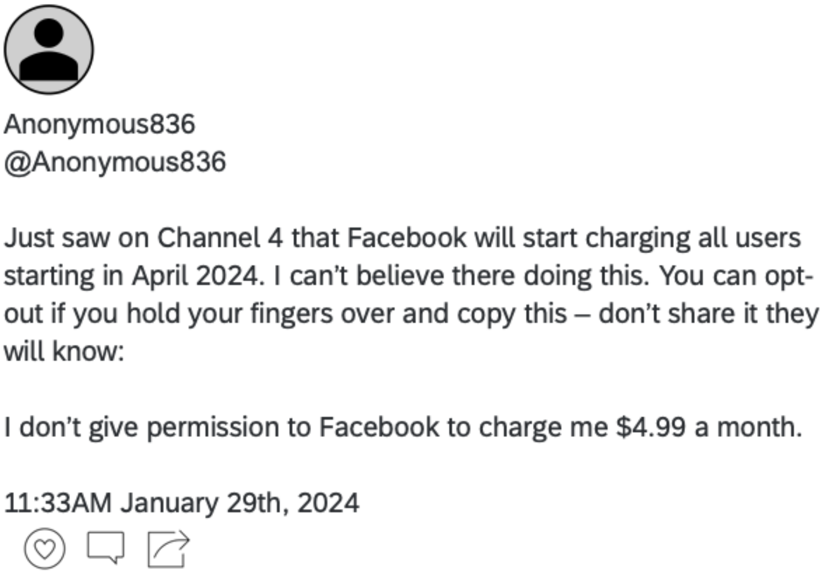
Appendix F

Chapter 4: OMEN Survey Posts

This appendix includes all the survey posts used in the survey experiment in Chapter 4.

F.1 Misinformation Detection Posts

These posts were used in the misinformation detection section of the survey experiment. Posts labeled as “misinformation” indicate that they are considered false or partially false and misleading by mainstream fact-checking organizations. Posts labeled as “real news” are from reputable sources or accounts posting reputable links and are considered true.



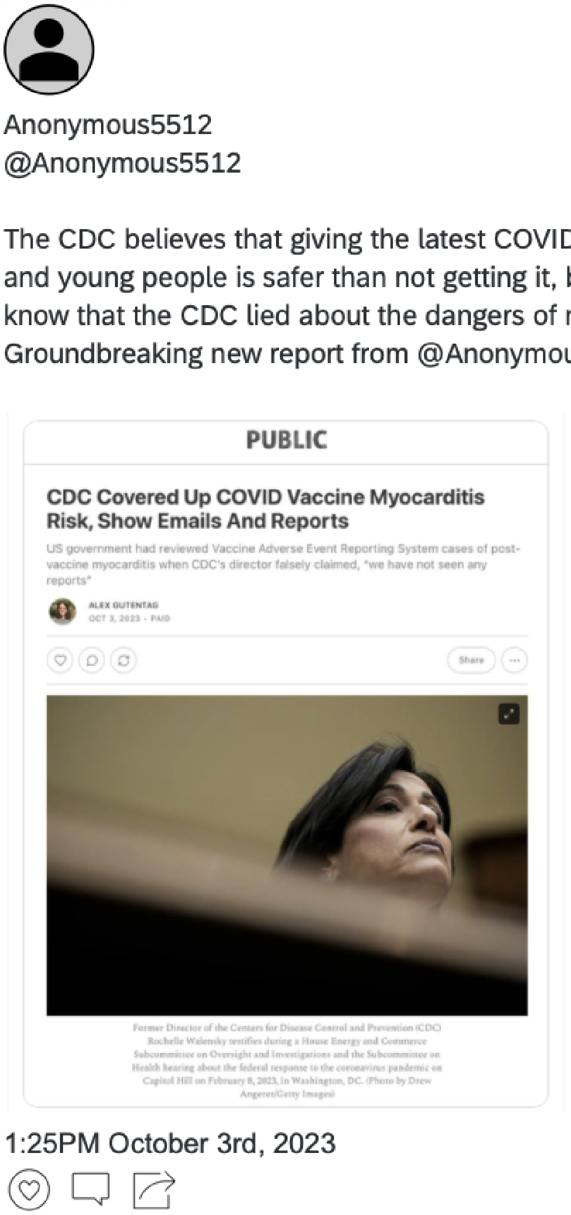
Anonymous836
@Anonymous836

Just saw on Channel 4 that Facebook will start charging all users starting in April 2024. I can’t believe there doing this. You can opt-out if you hold your fingers over and copy this – don’t share it they will know:

I don’t give permission to Facebook to charge me $4.99 a month.

11:33AM January 29th, 2024

Figure F.1: Pre-Test Misinformation Post #1.



Anonymous5512
@Anonymous5512

The CDC believes that giving the latest COVID19 booster to kids and young people is safer than not getting it, but it is not. Now we know that the CDC lied about the dangers of myocarditis. Groundbreaking new report from @Anonymous247.

PUBLIC

CDC Covered Up COVID Vaccine Myocarditis Risk, Show Emails And Reports

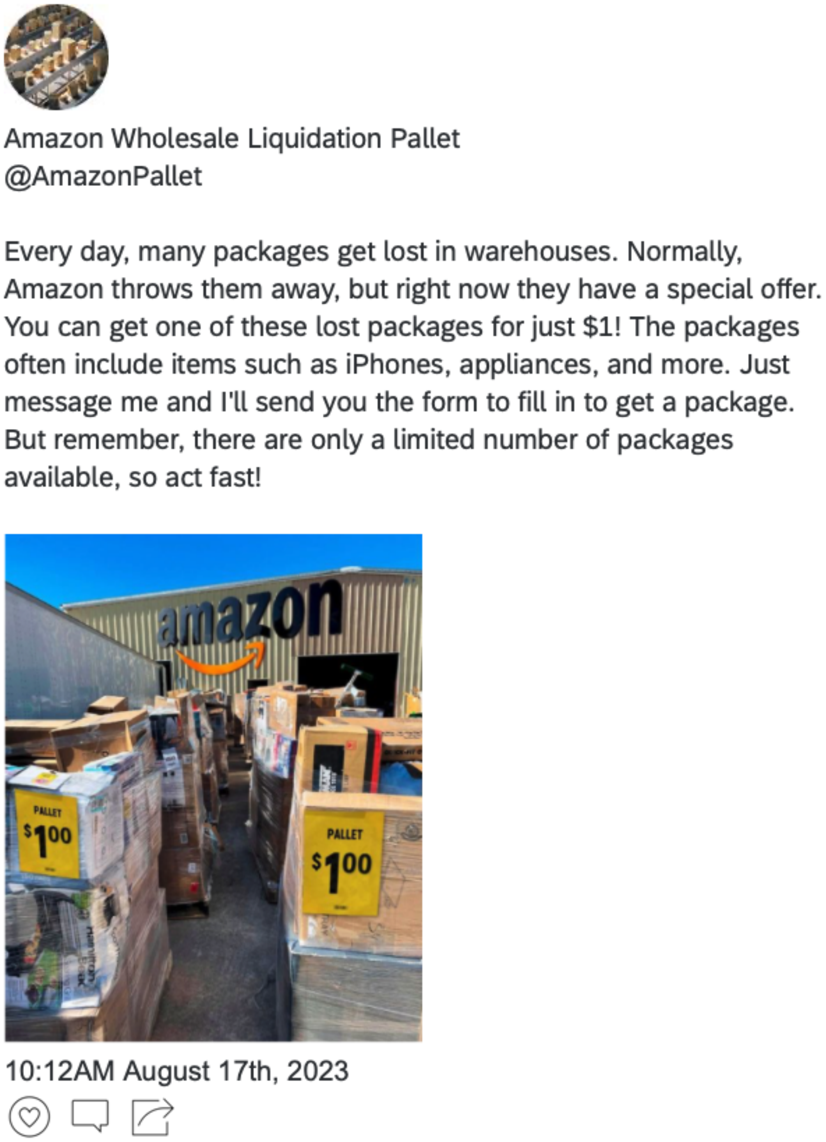
US government had reviewed Vaccine Adverse Event Reporting System cases of post-vaccine myocarditis when CDC's director falsely claimed, "we have not seen any reports"

ALEX GUTENTAG
OCT 3, 2023 · PAID

Former Director of the Centers for Disease Control and Prevention (CDC) Rochelle Walensky testifies during a House Energy and Commerce Subcommittee on Oversight and Investigations and the Subcommittee on Health hearing about the federal response to the coronavirus pandemic on Capitol Hill on February 8, 2023, in Washington, DC. (Photo by Drew Angerer/Getty Images)

1:25PM October 3rd, 2023

Figure F.2: Pre-Test Misinformation Post #2.

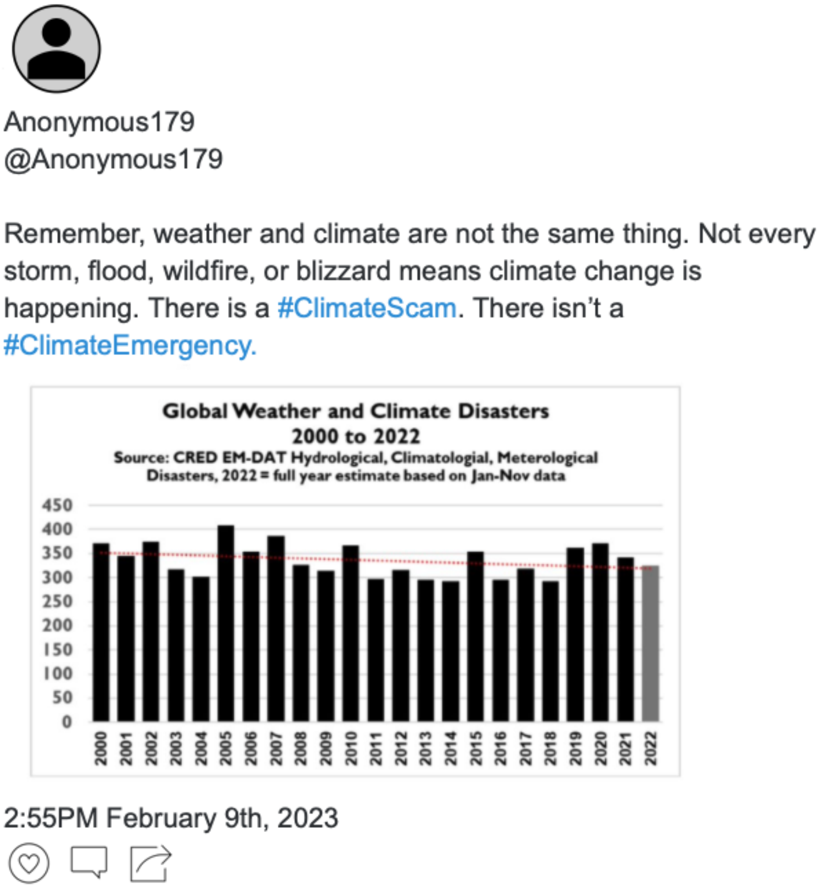


Amazon Wholesale Liquidation Pallet
@AmazonPallet

Every day, many packages get lost in warehouses. Normally, Amazon throws them away, but right now they have a special offer. You can get one of these lost packages for just $1! The packages often include items such as iPhones, appliances, and more. Just message me and I'll send you the form to fill in to get a package. But remember, there are only a limited number of packages available, so act fast!

10:12AM August 17th, 2023

Figure F.3: Pre-Test Misinformation Post #3.



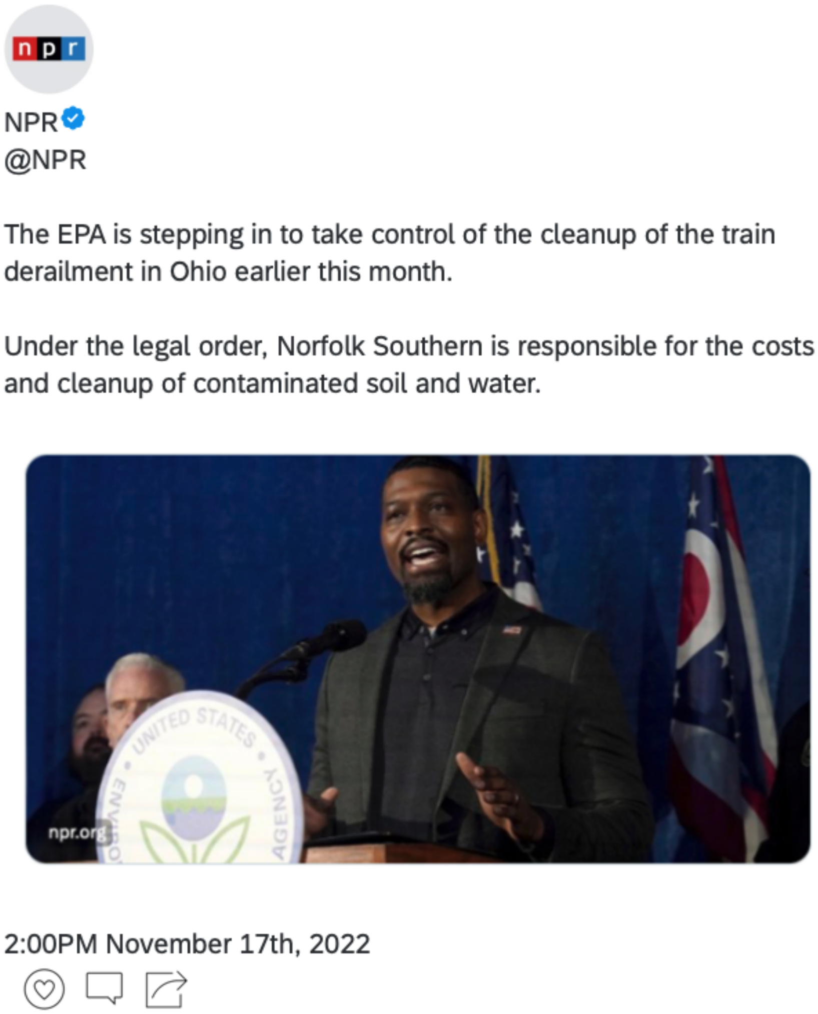
Anonymous179
@Anonymous179

Remember, weather and climate are not the same thing. Not every storm, flood, wildfire, or blizzard means climate change is happening. There is a #ClimateScam. There isn't a #ClimateEmergency.

Global Weather and Climate Disasters
2000 to 2022
Source: CRED EM-DAT Hydrological, Climatologial, Meterological Disasters, 2022 = full year estimate based on Jan-Nov data

2:55PM February 9th, 2023

Figure F.4: Pre-Test Misinformation Post #4.



NPR

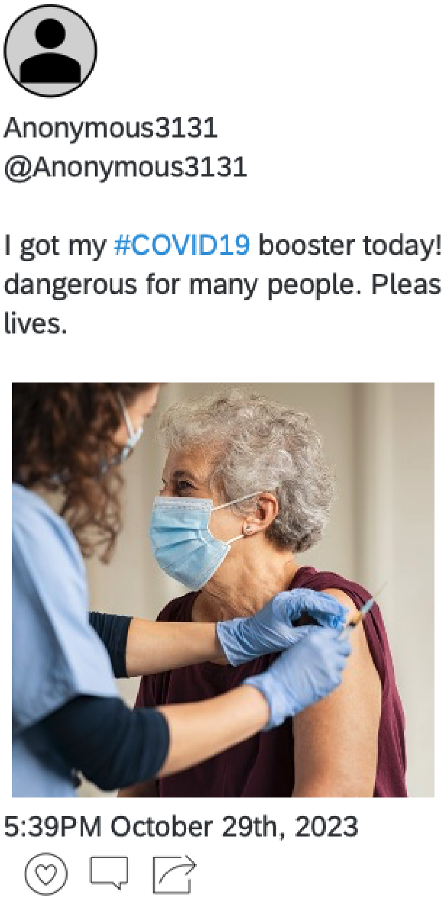
@NPR

The EPA is stepping in to take control of the cleanup of the train derailment in Ohio earlier this month.

Under the legal order, Norfolk Southern is responsible for the costs and cleanup of contaminated soil and water.

2:00PM November 17th, 2022

Figure F.5: Pre-Test Real News Post #1.

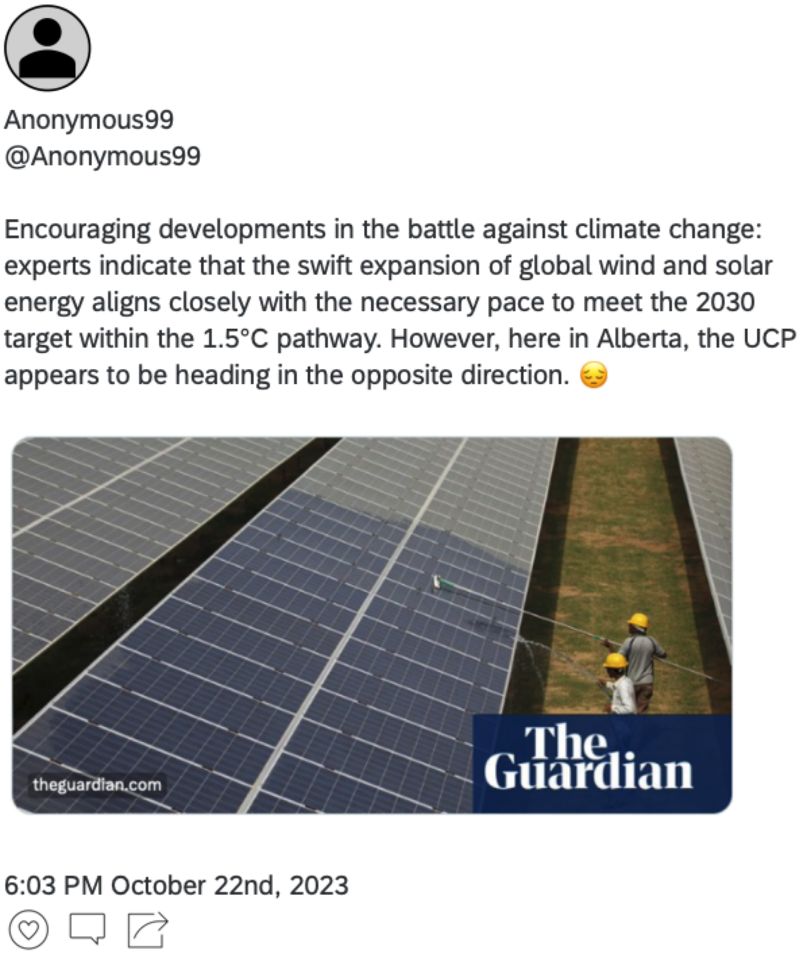


Anonymous3131
@Anonymous3131

I got my #COVID19 booster today! #Covid is still here and it's dangerous for many people. Please get vaccinated to help save lives.

5:39PM October 29th, 2023

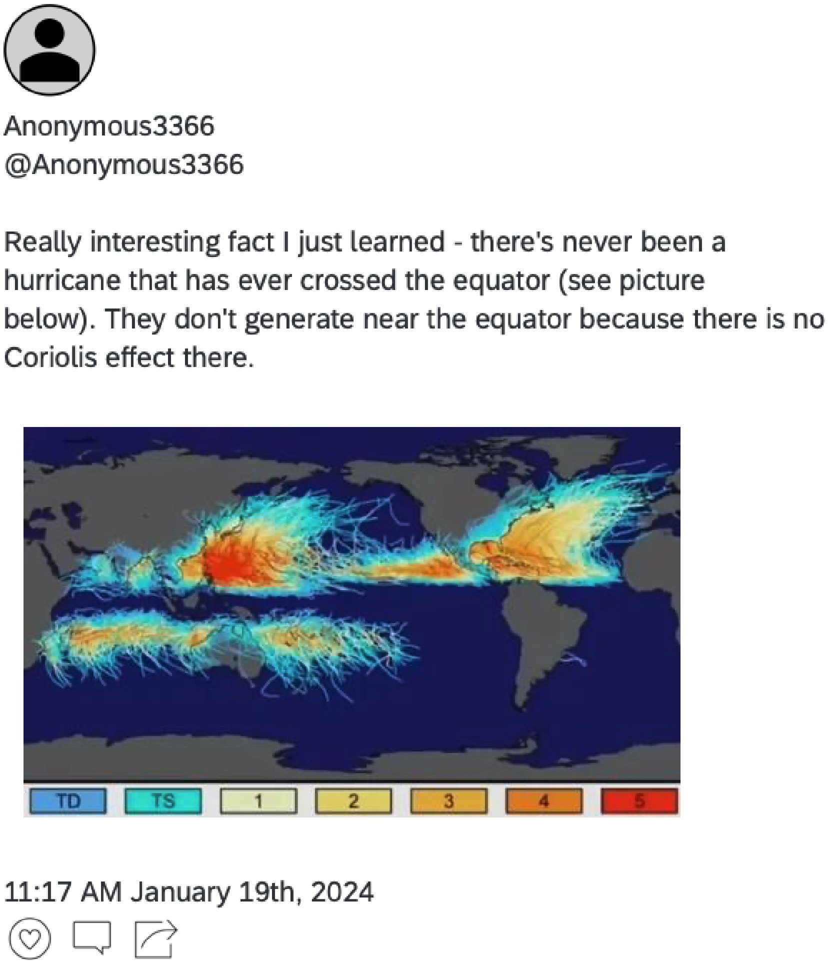
Figure F.6: Pre-Test Real News Post #2.


Anonymous99
@Anonymous99

Encouraging developments in the battle against climate change: experts indicate that the swift expansion of global wind and solar energy aligns closely with the necessary pace to meet the 2030 target within the 1.5°C pathway. However, here in Alberta, the UCP appears to be heading in the opposite direction. 😔

6:03 PM October 22nd, 2023

Figure F.7: Pre-Test Real News Post #3.

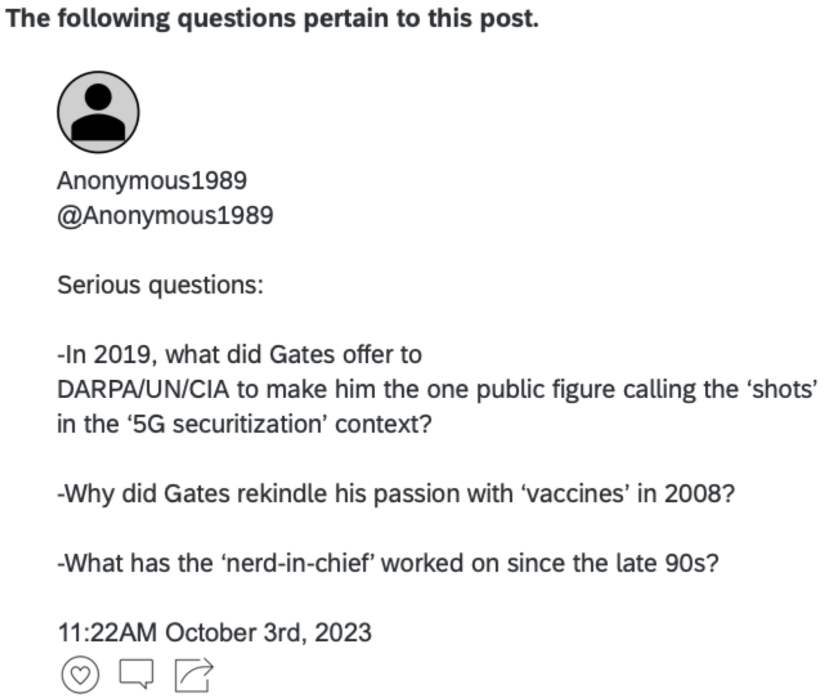


Anonymous3366
@Anonymous3366

Really interesting fact I just learned - there's never been a hurricane that has ever crossed the equator (see picture below). They don't generate near the equator because there is no Coriolis effect there.

11:17 AM January 19th, 2024

Figure F.8: Pre-Test Real News Post #4.



The following questions pertain to this post.

Anonymous1989
@Anonymous1989

Serious questions:

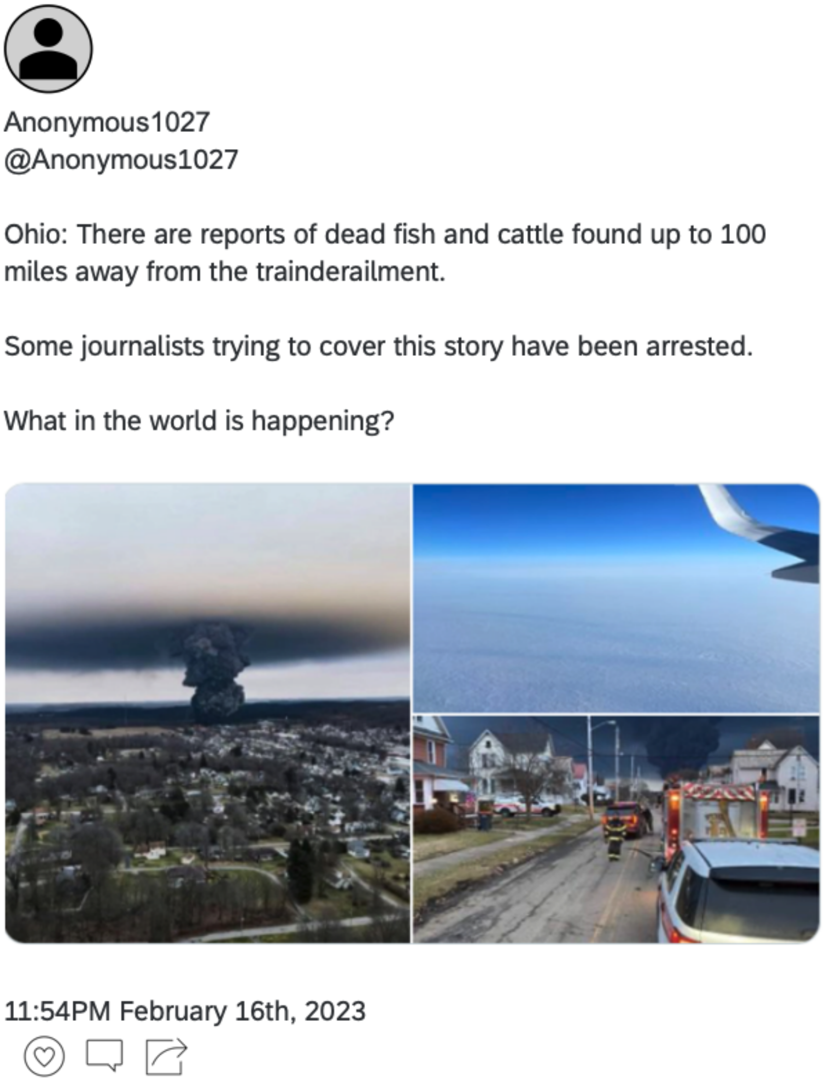
-In 2019, what did Gates offer to
DARPA/UN/CIA to make him the one public figure calling the ‘shots’ in the ‘5G securitization’ context?

-Why did Gates rekindle his passion with ‘vaccines’ in 2008?

-What has the ‘nerd-in-chief’ worked on since the late 90s?

11:22AM October 3rd, 2023

Figure F.9: Post-Test Misinformation Post #1.



Anonymous1027
@Anonymous1027

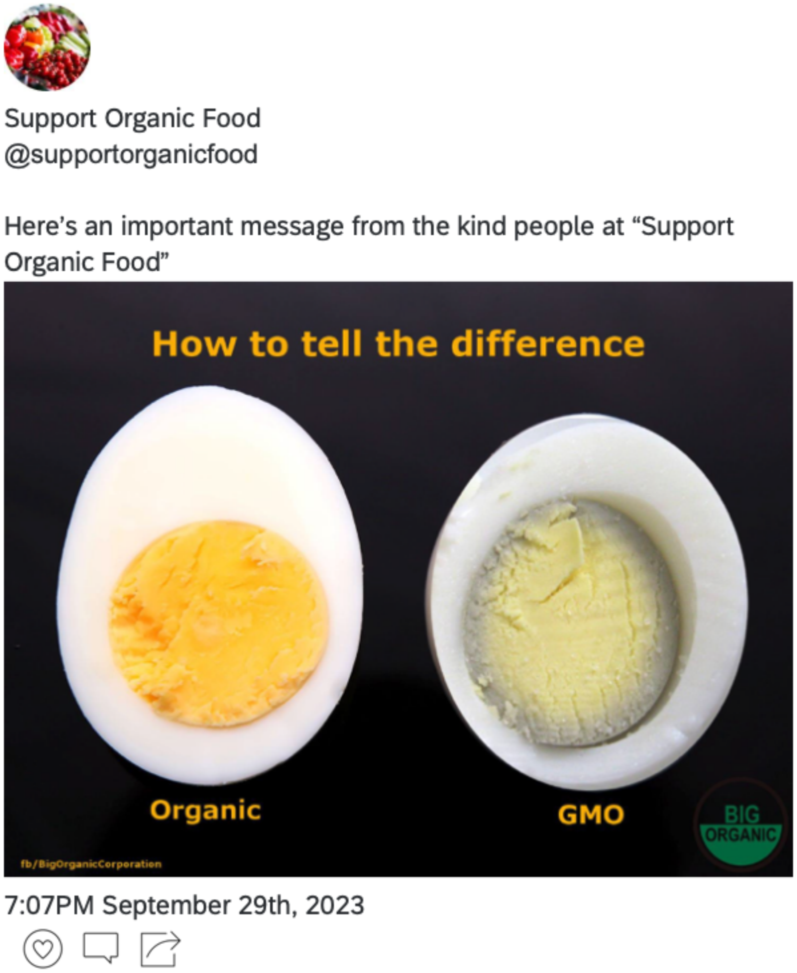
Ohio: There are reports of dead fish and cattle found up to 100 miles away from the trainderailment.

Some journalists trying to cover this story have been arrested.

What in the world is happening?

11:54PM February 16th, 2023

Figure F.10: Post-Test Misinformation Post #2.

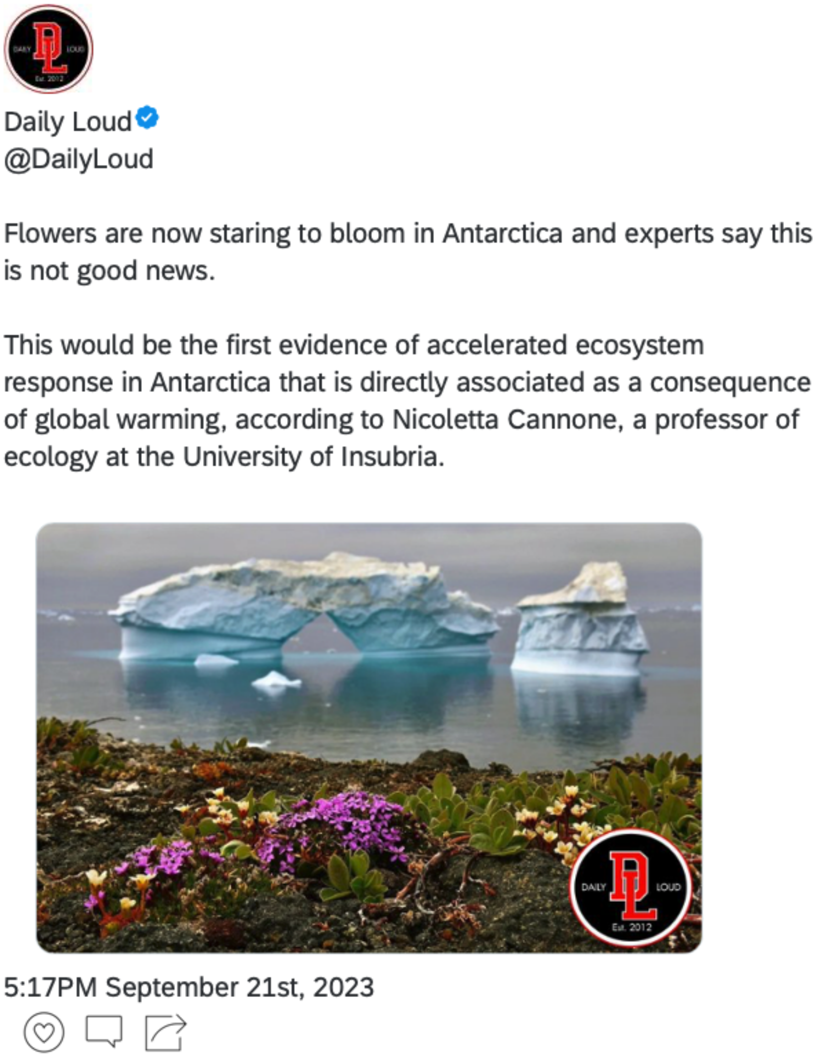


Support Organic Food
@supportorganicfood

Here's an important message from the kind people at "Support Organic Food"

7:07PM September 29th, 2023

Figure F.11: Post-Test Misinformation Post #3.



Daily Loud

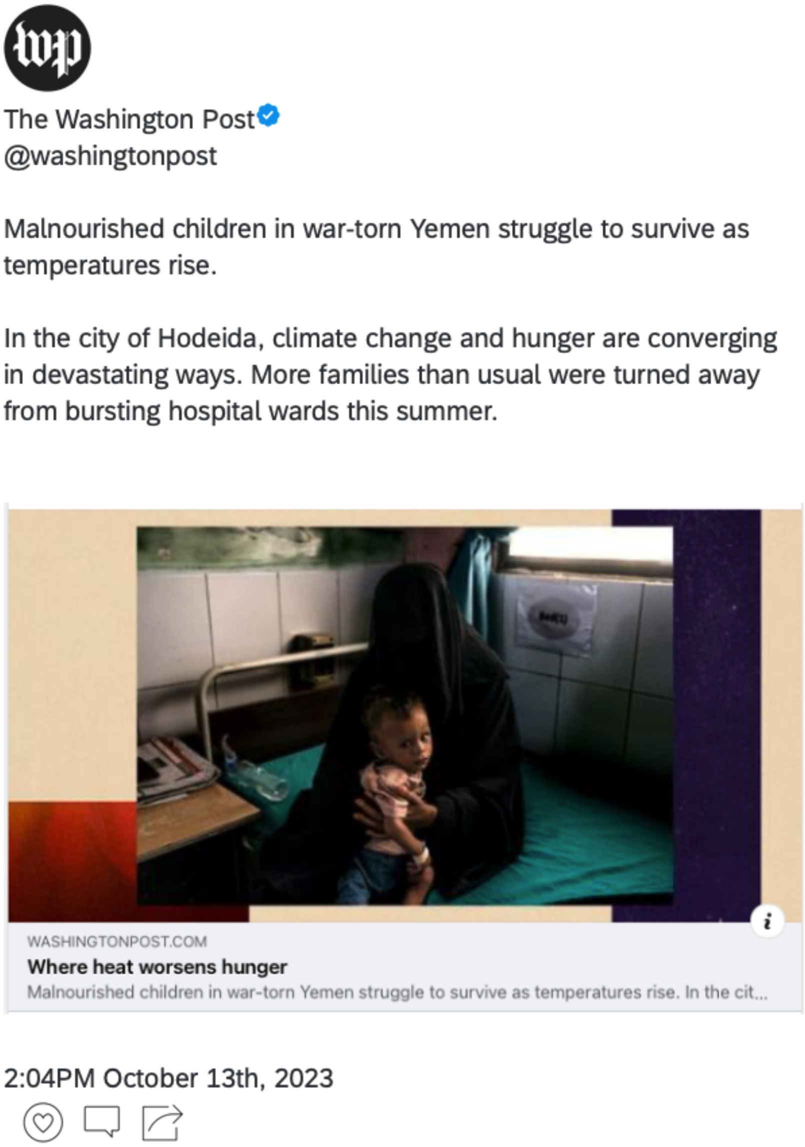
@DailyLoud

Flowers are now staring to bloom in Antarctica and experts say this is not good news.

This would be the first evidence of accelerated ecosystem response in Antarctica that is directly associated as a consequence of global warming, according to Nicoletta Cannone, a professor of ecology at the University of Insubria.

5:17PM September 21st, 2023

Figure F.12: Post-Test Misinformation Post #4.



The Washington Post
@washingtonpost

Malnourished children in war-torn Yemen struggle to survive as temperatures rise.

In the city of Hodeida, climate change and hunger are converging in devastating ways. More families than usual were turned away from bursting hospital wards this summer.

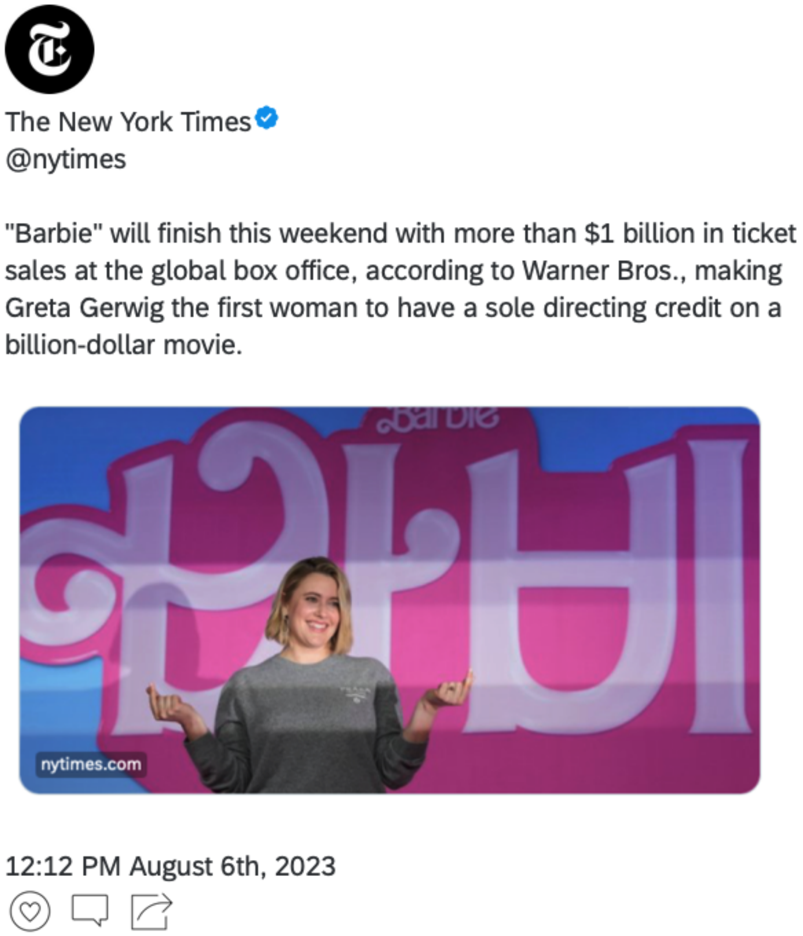
WASHINGTONPOST.COM

Where heat worsens hunger

Malnourished children in war-torn Yemen struggle to survive as temperatures rise. In the cit...

2:04PM October 13th, 2023

Figure F.13: Post-Test Real News Post #1.

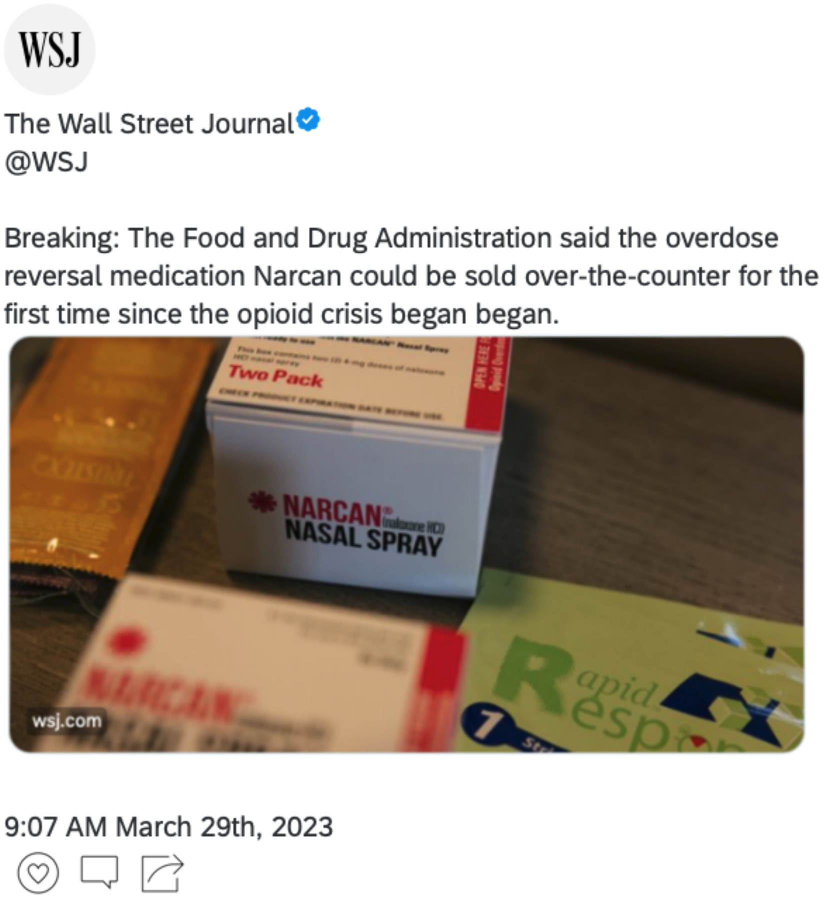


The New York Times
@nytimes

"Barbie" will finish this weekend with more than $1 billion in ticket sales at the global box office, according to Warner Bros., making Greta Gerwig the first woman to have a sole directing credit on a billion-dollar movie.

12:12 PM August 6th, 2023

Figure F.14: Post-Test Real News Post #2.

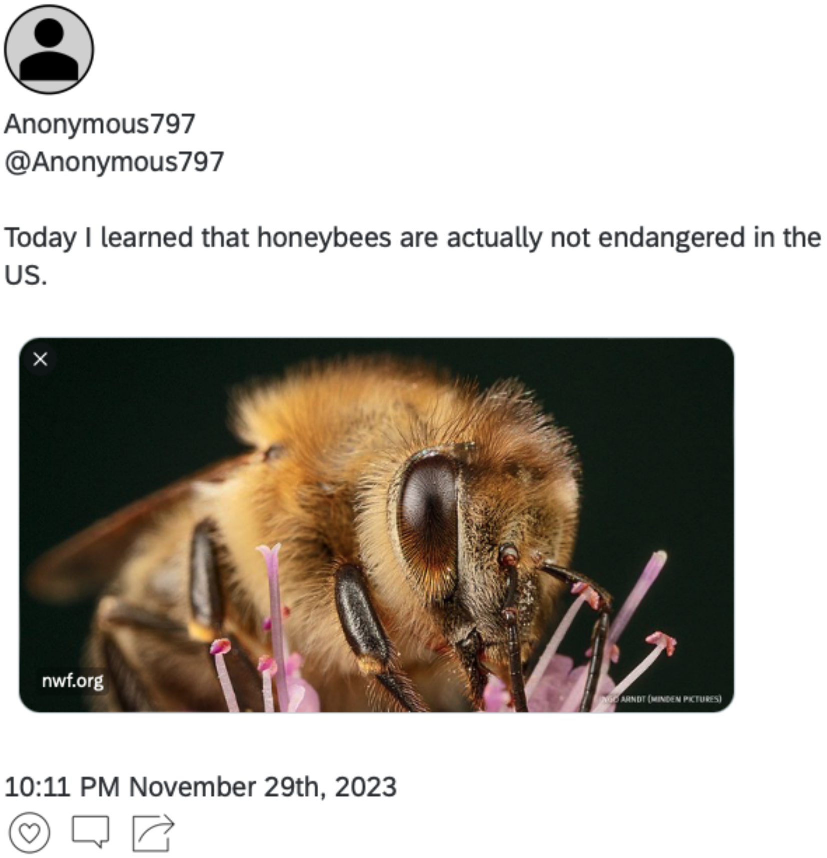


The Wall Street Journal
@WSJ

Breaking: The Food and Drug Administration said the overdose reversal medication Narcan could be sold over-the-counter for the first time since the opioid crisis began began.

9:07 AM March 29th, 2023

Figure F.15: Post-Test Real News Post #3.



Anonymous797
@Anonymous797

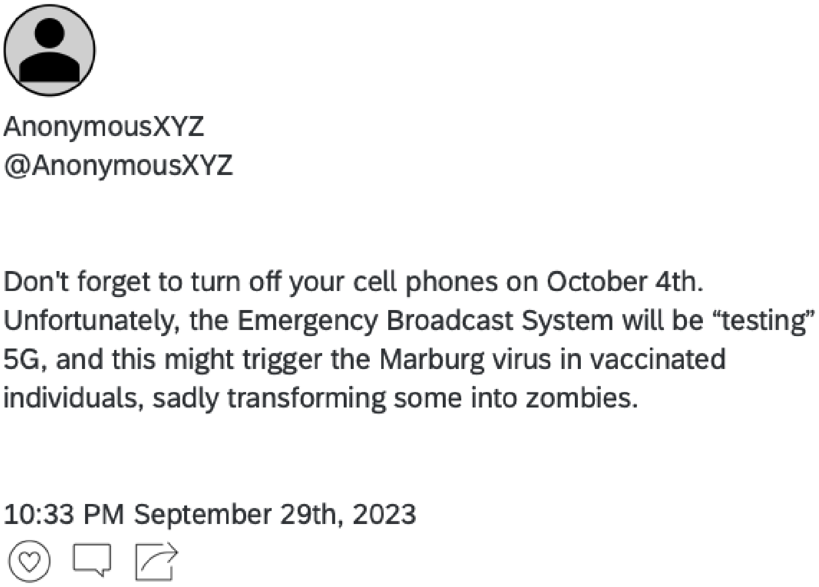
Today I learned that honeybees are actually not endangered in the US.

10:11 PM November 29th, 2023

Figure F.16: Post-Test Real News Post #4.

F.2 Countering Posts

These posts were used in the countering section of the survey experiment. All posts were described as being false to participants and were selected so as to be uncontroversially and obviously considered false or “misinformation” by all participants.

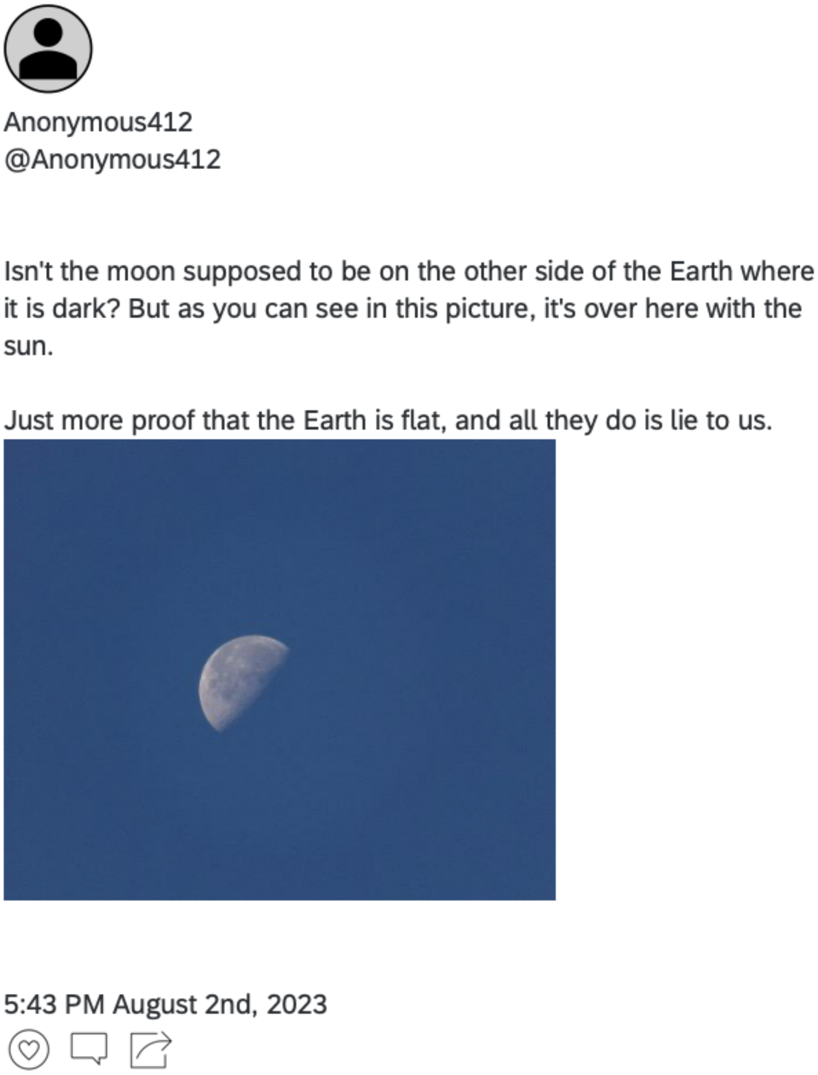


AnonymousXYZ
@AnonymousXYZ

Don't forget to turn off your cell phones on October 4th. Unfortunately, the Emergency Broadcast System will be “testing” 5G, and this might trigger the Marburg virus in vaccinated individuals, sadly transforming some into zombies.

10:33 PM September 29th, 2023

Figure F.17: Pre-Test Countering Post #1.



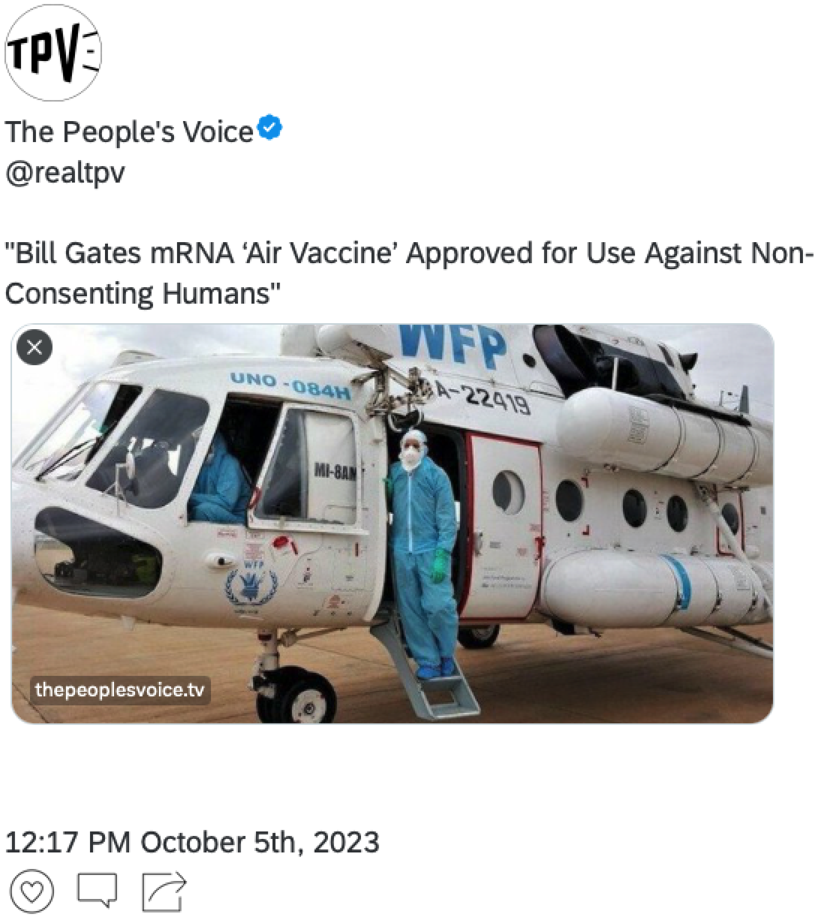
Anonymous412
@Anonymous412

Isn't the moon supposed to be on the other side of the Earth where it is dark? But as you can see in this picture, it's over here with the sun.

Just more proof that the Earth is flat, and all they do is lie to us.

5:43 PM August 2nd, 2023

Figure F.18: Pre-Test Countering Post #2.



TPV

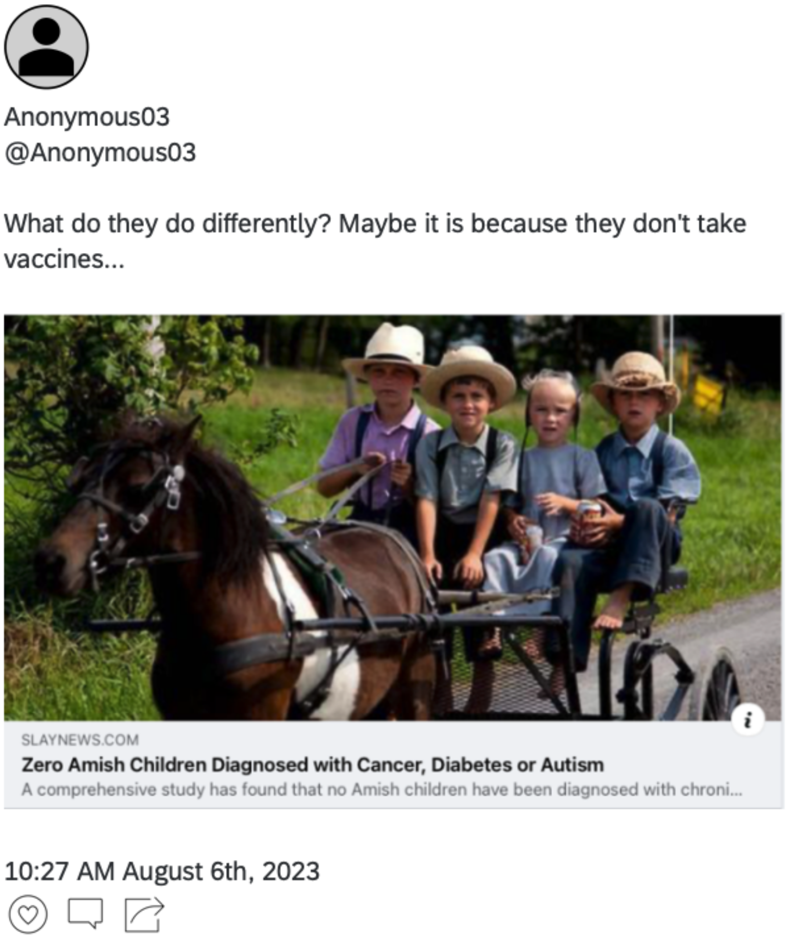
The People's Voice

@realtpv

"Bill Gates mRNA ‘Air Vaccine’ Approved for Use Against Non-Consenting Humans"

12:17 PM October 5th, 2023

Figure F.19: Pre-Test Countering Post #3.



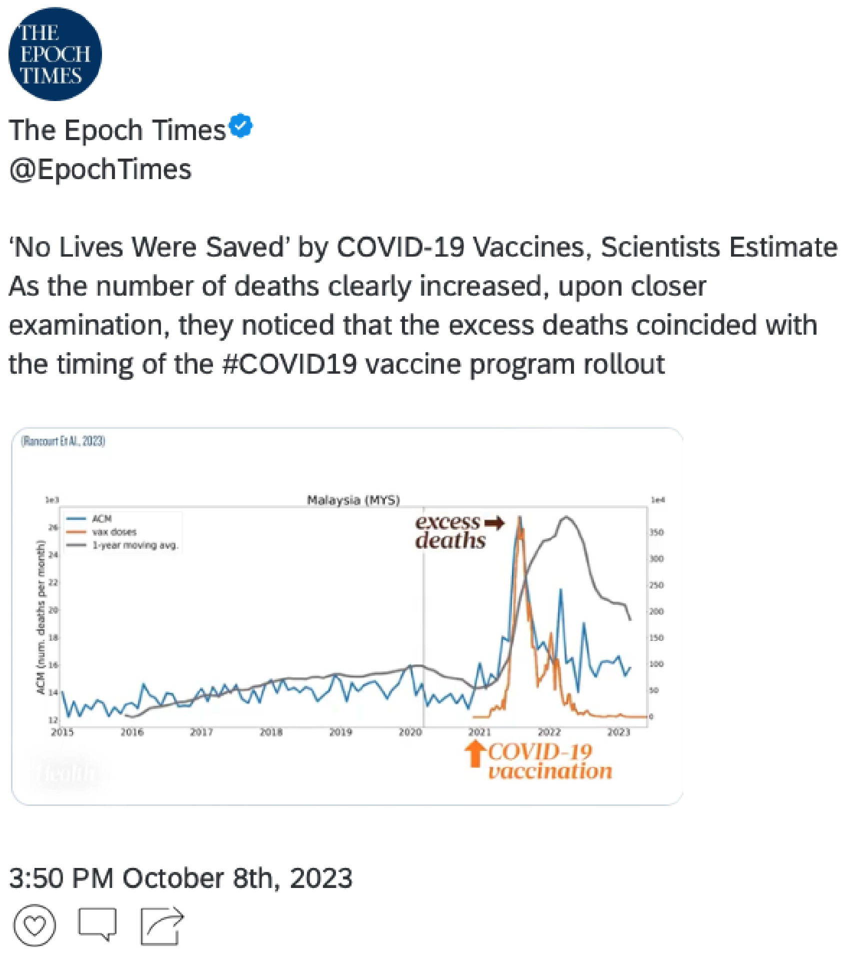
Anonymous03
@Anonymous03

What do they do differently? Maybe it is because they don't take vaccines...

SLAYNEWS.COM
Zero Amish Children Diagnosed with Cancer, Diabetes or Autism
A comprehensive study has found that no Amish children have been diagnosed with chroni...

10:27 AM August 6th, 2023

Figure F.20: Pre-Test Countering Post #4.



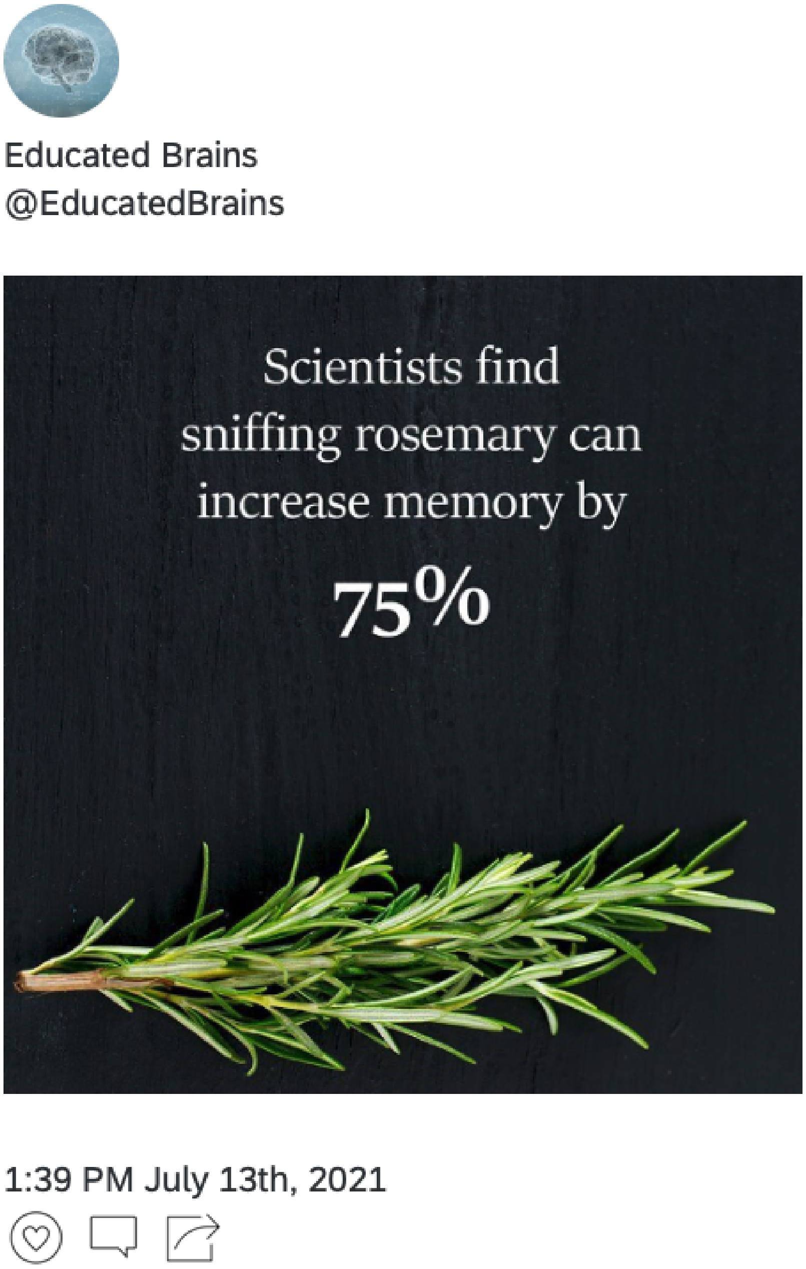
The Epoch Times
@EpochTimes

‘No Lives Were Saved’ by COVID-19 Vaccines, Scientists Estimate
As the number of deaths clearly increased, upon closer examination, they noticed that the excess deaths coincided with the timing of the #COVID19 vaccine program rollout

(Rancourt Et Al., 2023)

3:50 PM October 8th, 2023

Figure F.21: Post-Test Countering Post #1.



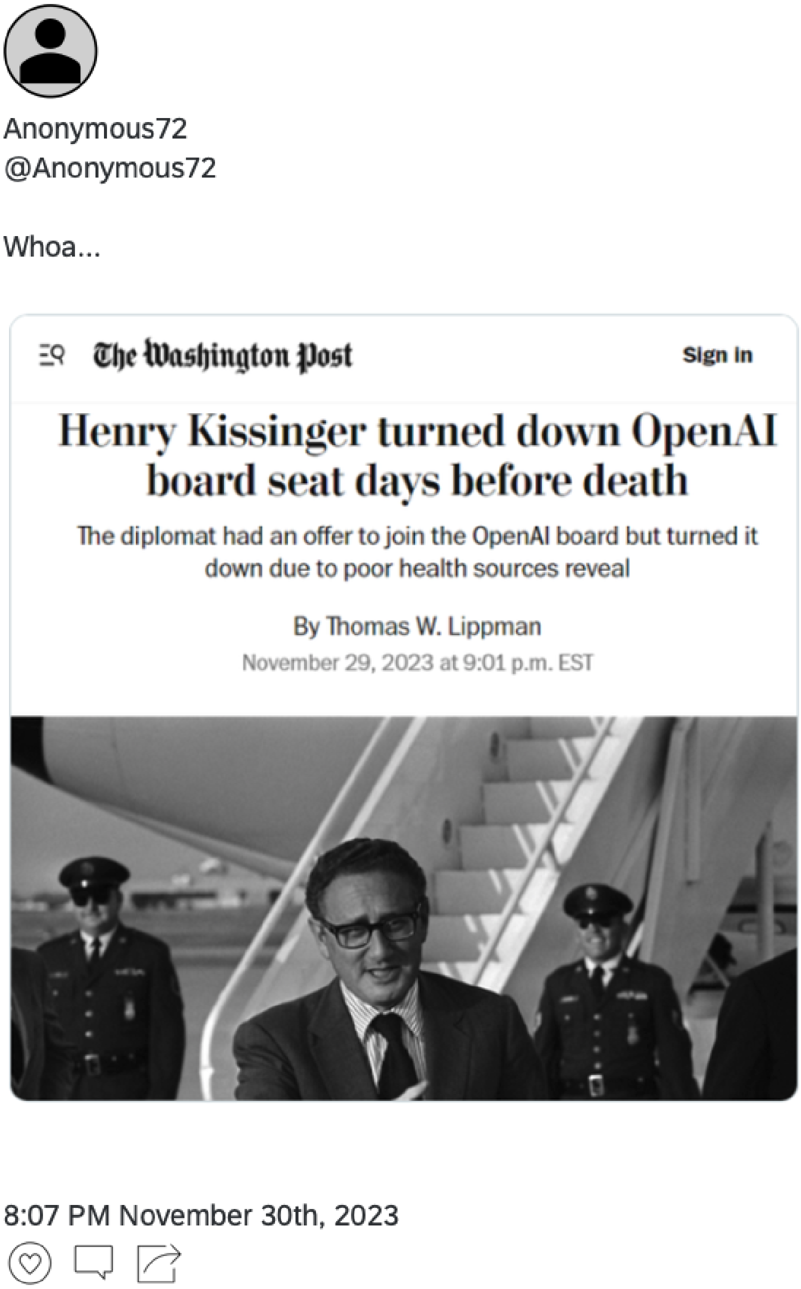
Educated Brains
@EducatedBrains

Scientists find sniffing rosemary can increase memory by

75%

1:39 PM July 13th, 2021

Figure F.22: Post-Test Countering Post #2.



Anonymous72
@Anonymous72

Whoa...

The Washington Post

Sign in

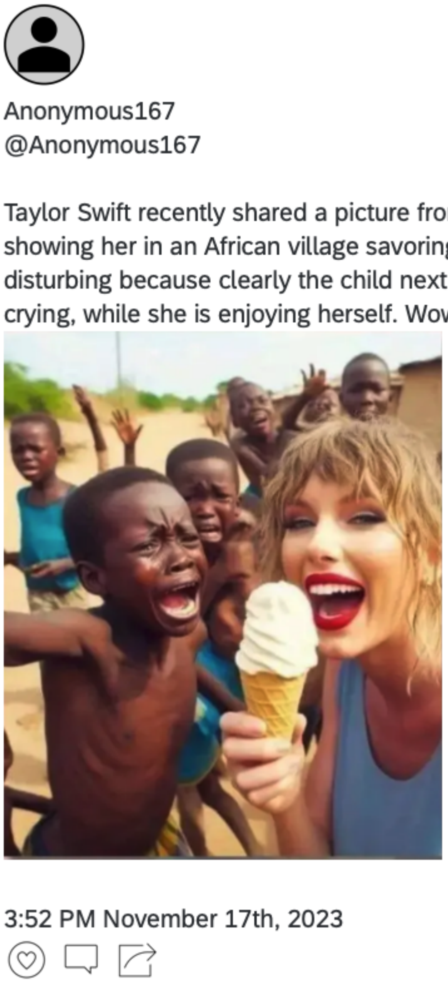
Henry Kissinger turned down OpenAI board seat days before death

The diplomat had an offer to join the OpenAI board but turned it down due to poor health sources reveal

By Thomas W. Lippman
November 29, 2023 at 9:01 p.m. EST

8:07 PM November 30th, 2023

Figure F.23: Post-Test Countering Post #3.



Anonymous167
@Anonymous167

Taylor Swift recently shared a picture from the latest stop on her tour, showing her in an African village savoring some ice cream. This picture is so disturbing because clearly the child next to her is extremely distressed and crying, while she is enjoying herself. Wow....

3:52 PM November 17th, 2023

Figure F.24: Post-Test Countering Post #4.

Appendix G

Chapter 5: Robustness Analysis

In the main analysis of individual differences in support and perceived attributes of misinformation interventions (see Chapter 5 Table 5.3), we ran OLS regressions to predict average support, perceived fairness, perceived effectiveness, and perceived intrusiveness of interventions as a function of various demographic variables. To supplement this analysis, we varied the partisanship groups to show that the results were robust to different partisan category mappings. Table G.1 separates out the "Independent" and "Other/unaffiliated" categories rather than combining them. Table G.2 maps partisan categories to corresponding numeric values (strong Democrat to strong Republican categories mapped to numeric values 0 to 4). The results are substantially similar to the primary analysis.

Based on these results and the primary analysis in RQ3, we identified partisanship, political ideology, and gender for further examination. In the main ad hoc analysis, we examined the impact of adding partisanship and gender to the main regression model to determine if these factors interacted with the implementer or the perceived attributes of the interventions to predict support (see Table 5.5). Due to the high correlation between partisanship and political ideology (see Table 5.4), our primary ad hoc analysis included only partisanship and gender. To supplement this analysis, we conducted the same model with political ideology included instead of partisanship (see Table G.3). The results are substantially similar to the primary analysis, with more conservative participants placing a higher importance on fairness than liberal participants.

Table G.1: OLS regressions predicting average support, perceived fairness, perceived effectiveness, and perceived intrusiveness of interventions as a function of demographic variables. Same analysis as Table 5.3, except partisanship separated out the “Other” party from “Independent”.

	Dependent variable:			
	Avg. support rating	Avg. fairness rating	Avg. effectiveness rating	Avg. intrusiveness rating
Age	0.020	0.044*	−0.023	0.038*
	(0.021)	(0.021)	(0.019)	(0.018)
Male : Female	−0.201***	−0.139*	−0.217***	0.103*
	(0.054)	(0.054)	(0.050)	(0.048)
Other (e.g., non-binary) : Female	−0.074	−0.252	−0.271	0.105
	(0.175)	(0.178)	(0.162)	(0.156)
Education	0.013	0.003	0.007	0.016
	(0.022)	(0.022)	(0.020)	(0.019)
Income	0.032*	0.046**	0.012	−0.005
	(0.016)	(0.016)	(0.015)	(0.014)
Independent : Democrat	−0.260***	−0.243**	−0.185**	0.191**
	(0.075)	(0.076)	(0.070)	(0.067)
Other party : Democrat	−0.576**	−0.596**	−0.620***	0.461**
	(0.199)	(0.202)	(0.185)	(0.178)
Republican : Democrat	−0.275*	−0.281**	−0.193	0.152
	(0.107)	(0.109)	(0.099)	(0.095)
Political Ideology	−0.300***	−0.268***	−0.190***	0.134***
	(0.037)	(0.037)	(0.034)	(0.033)
Misinformation Exposure	−0.015	−0.008	−0.048	0.039
	(0.027)	(0.027)	(0.025)	(0.024)
Constant	4.240***	4.042***	4.002***	2.493***
	(0.122)	(0.124)	(0.114)	(0.109)
Observations	1,010	1,010	1,010	1,010
R^2	0.246	0.208	0.156	0.097
Adjusted R^2	0.238	0.200	0.147	0.088
Residual Std. Error (df = 999)	0.833	0.846	0.773	0.743
F Statistic (df = 10; 999)	32.525***	26.264***	18.443***	10.782***
Note:				*p<0.05; **p<0.01; ***p<0.001

Table G.2: OLS regressions predicting average support, perceived fairness, perceived effectiveness, and perceived intrusiveness of interventions as a function of demographic variables. Same analysis as Table 5.3, except partisanship was mapped to numeric values 0 to 4, with strong Democrats mapped to 0 and strong Republicans mapped to 4.

	Dependent variable:			
	Avg. support rating	Avg. fairness rating	Avg. effectiveness rating	Avg. intrusiveness rating
Age	0.018	0.043*	−0.024	0.039*
	(0.021)	(0.021)	(0.019)	(0.019)
Male : Female	−0.199***	−0.136*	−0.214***	0.103*
	(0.054)	(0.054)	(0.050)	(0.048)
Other (e.g., non-binary) : Female	−0.130	−0.306	−0.338*	0.158
	(0.173)	(0.175)	(0.161)	(0.154)
Education	0.014	0.003	0.007	0.015
	(0.022)	(0.022)	(0.020)	(0.019)
Income	0.031*	0.045**	0.012	−0.005
	(0.016)	(0.016)	(0.015)	(0.014)
Partisanship	−0.128***	−0.135***	−0.098**	0.080**
	(0.033)	(0.034)	(0.031)	(0.030)
Political Ideology	−0.263***	−0.227***	−0.157***	0.108**
	(0.038)	(0.038)	(0.035)	(0.034)
Misinformation Exposure	−0.015	−0.008	−0.048	0.040
	(0.027)	(0.027)	(0.025)	(0.024)
Constant	4.246***	4.054***	4.001***	2.503***
	(0.120)	(0.122)	(0.112)	(0.108)
Observations	1,010	1,010	1,010	1,010
R^2	0.244	0.208	0.151	0.092
Adjusted R^2	0.238	0.202	0.144	0.085
Residual Std. Error (df = 1001)	0.833	0.846	0.775	0.745
F Statistic (df = 8; 1001)	40.316***	32.834***	22.280***	12.742***
Note:				*p<0.05; **p<0.01; ***p<0.001

Table G.3: Adhoc analysis: OLS regression predicting support for a misinformation intervention as a function of perceived fairness, perceived effectiveness, perceived intrusiveness, implementer (reference level: social media company), gender (reference level: Female), and ideology (very liberal to very conservative categories mapped to numeric values 0 to 4) with robust standard errors clustered on participant and intervention.

	Dependent variable: Support	
	Estimate	Std. Err.
Implementer (platform : government)	0.159	(0.099)
Perceived fairness	0.539^{***}	(0.029)
Perceived effectiveness	0.329^{***}	(0.025)
Perceived intrusiveness	-0.073^{***}	(0.017)
Gender (male : female)	-0.218^{*}	(0.088)
Gender (other: female)	0.634^{*}	(0.287)
Ideology	-0.128^{***}	(0.038)
Implementer x Perceived fairness	-0.076^{**}	(0.024)
Implementer x Perceived effectiveness	0.063^{**}	(0.022)
Implementer x Perceived intrusiveness	−0.034	(0.014)
Implementer x Gender (male : female)	0.003	(0.030)
Implementer x Gender (other: female)	0.171	(0.104)
Implementer x Ideology	−0.016	(0.014)
Perceived fairness x Gender (male : female)	0.039	(0.023)
Perceived fairness x Gender (other : female)	−0.086	(0.059)
Perceived fairness x Ideology	**0.028****	(0.010)
Perceived effectiveness x Gender (male : female)	−0.007	(0.022)
Perceived effectiveness x Gender (other : female)	−0.033	(0.066)
Perceived effectiveness x Ideology	−0.016	(0.009)
Perceived intrusiveness x Gender (male : female)	0.018	(0.014)
Perceived intrusiveness x Gender (other : female)	−0.051	(0.047)
Perceived intrusiveness x Ideology	0.001	(0.006)
Constant	0.991^{***}	(0.114)
Observations	8071	
R^2	0.766	
Adjusted R^2	0.766	
Note:	$^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$	

Appendix H

Chapter 5: Demographic Analysis

H.1 Intervention Ratings by Partisanship

Table H.1 displays the average support ratings for each intervention by political party, while Table H.2 shows the 10 interventions ranked by the highest average support rating per party affiliation. Figures H.1-H.10 display the average Likert ratings along with the 95% confidence intervals for support and perceptions by political party and intervention implementer for each of the 10 interventions as defined in Table 5.1.

Table H.1: Average support ratings and standard deviations per intervention by political party. The top 3 interventions by party are bolded

Intervention	**Strong Democrat**	**Weak Democrat**	**Independent / other**	**Weak Republican**	**Strong Republican**
1 Temporary delay	2.93 (1.23)	2.71 (1.21)	2.48 (1.22)	2.43 (1.19)	2.06 (1.33)
2 Fact-check ads	4.47 (0.90)	**4.40 (0.91)**	3.72 (1.39)	3.60 (1.38)	**3.15 (1.52)**
3 De-emphasize	4.22 (0.98)	3.97 (1.12)	3.35 (1.38)	3.13 (1.28)	2.77 (1.44)
4 Remove posts	4.22 (1.04)	4.12 (1.12)	3.31 (1.52)	3.27 (1.56)	2.72 (1.55)
5 Permanent ban	3.87 (1.24)	3.53 (1.29)	3.16 (1.47)	2.60 (1.43)	2.55 (1.53)
6 Notify users	**4.50 (0.82)**	**4.36 (0.92)**	**3.74 (1.38)**	**3.72 (1.35)**	**3.27 (1.57)**
7 Public labels	**4.59 (0.73)**	**4.39 (0.89)**	**3.75 (1.33)**	**3.66 (1.32)**	2.97 (1.53)
8 Media literacy	**4.48 (0.84)**	4.18 (0.88)	**3.75 (1.20)**	**4.08 (1.05)**	**3.14 (1.36)**
9 Local media	3.60 (1.03)	3.37 (1.03)	3.25 (1.20)	3.30 (1.30)	3.01 (1.32)
10 Release data	4.41 (0.85)	4.14 (0.91)	3.66 (1.32)	3.60 (1.24)	3.13 (1.40)

Table H.2: Top-ranked items by average support rating for each political affiliation.

Strong Democrat	**Weak Democrat**	**Independent/ other**	**Weak Republican**	**Strong Republican**
7 Public labels	2 Fact-check ads	8 Media literacy	8 Media literacy	6 Notify users
6 Notify users	7 Public labels	7 Public labels	6 Notify users	2 Fact-check ads
8 Media literacy	6 Notify users	6 Notify users	7 Public labels	8 Media literacy
2 Fact-check ads	8 Media literacy	2 Fact-check ads	10 Release data	10 Release data
10 Release data	10 Release data	10 Release data	2 Fact-check ads	9 Local media
4 Remove posts	4 Remove posts	3 De-emphasize	9 Local media	7 Public labels
3 De-emphasize	3 De-emphasize	4 Remove posts	4 Remove posts	3 De-emphasize
5 Permanent ban	5 Permanent ban	9 Local media	3 De-emphasize	4 Remove posts
9 Local media	9 Local media	5 Permanent ban	5 Permanent ban	5 Permanent ban
1 Temporary delay	1 Temporary delay	1 Temporary delay	1 Temporary delay	1 Temporary delay

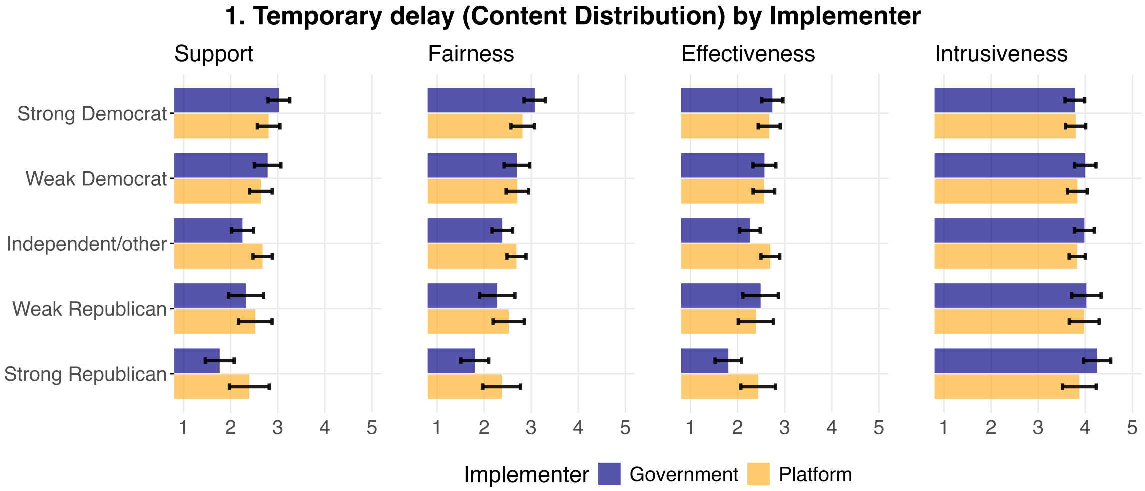


Figure H.1: Average Likert ratings for the **friction** intervention: temporarily delaying users from posting content they have not opened.

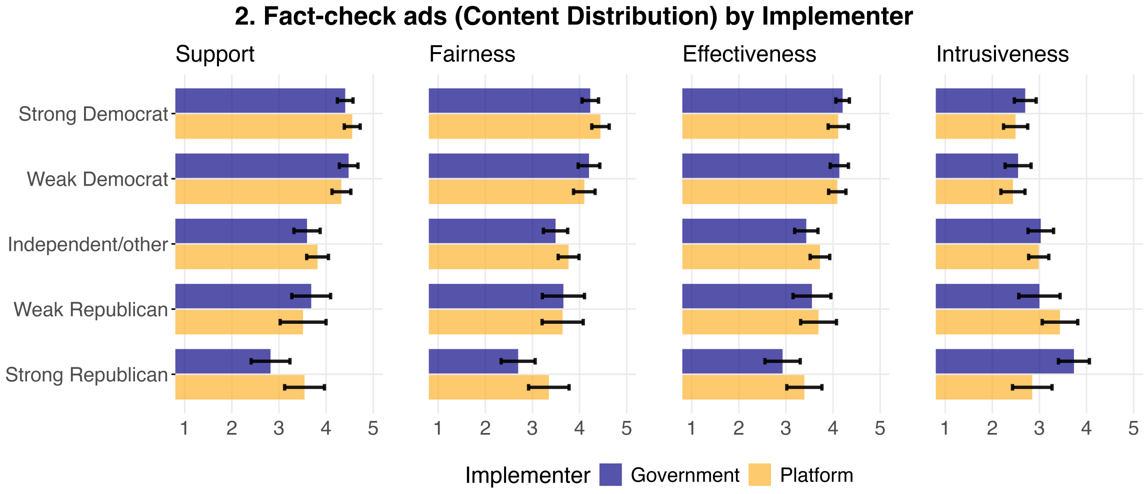


Figure H.2: Average Likert ratings for the **advertising policy** intervention: requiring all ads to be put through a fact-checking process.

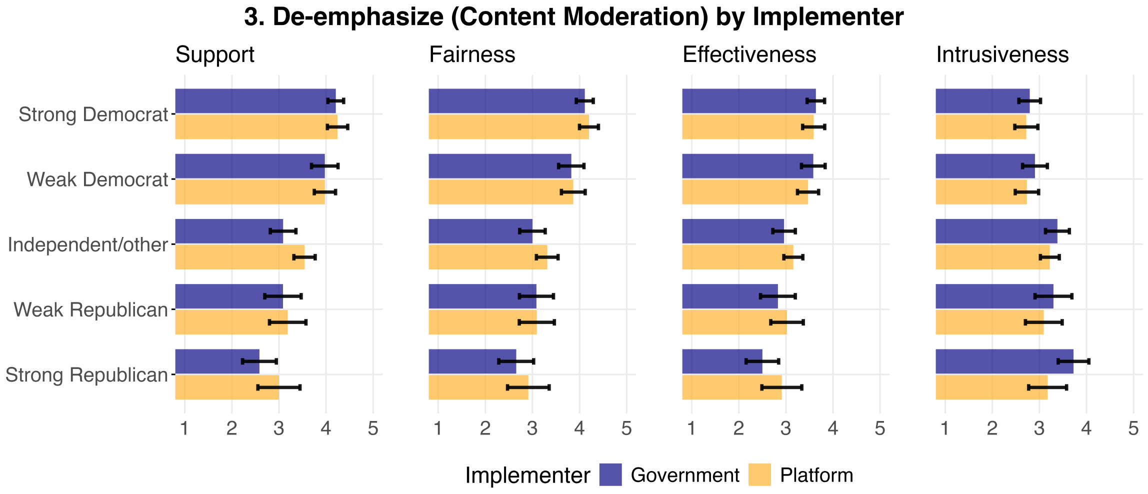


Figure H.3: Average Likert ratings for the **algorithmic downranking** intervention: de-emphasizing posts that are verified to contain misinformation.

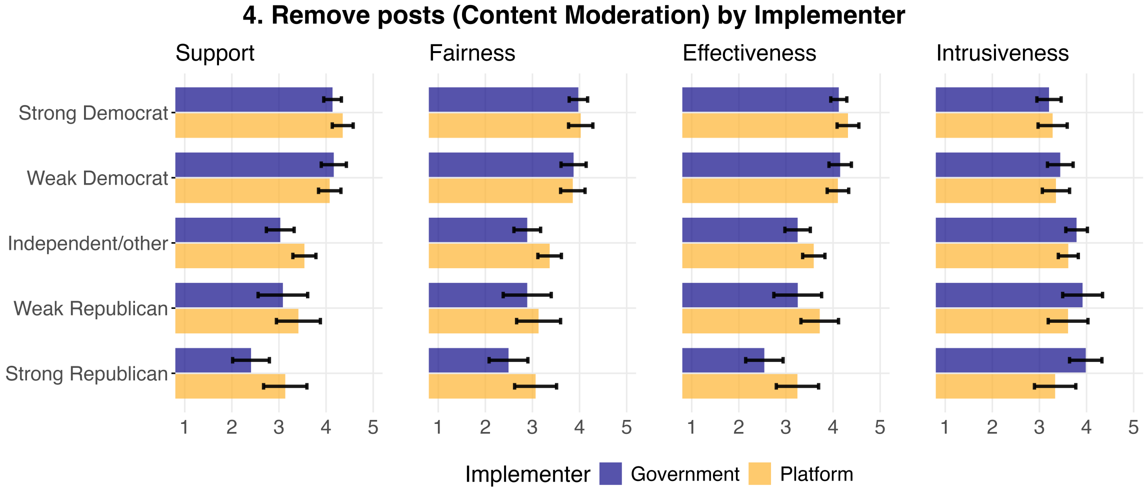


Figure H.4: Average Likert ratings for the **content removal** intervention: removing posts verified to contain misinformation.

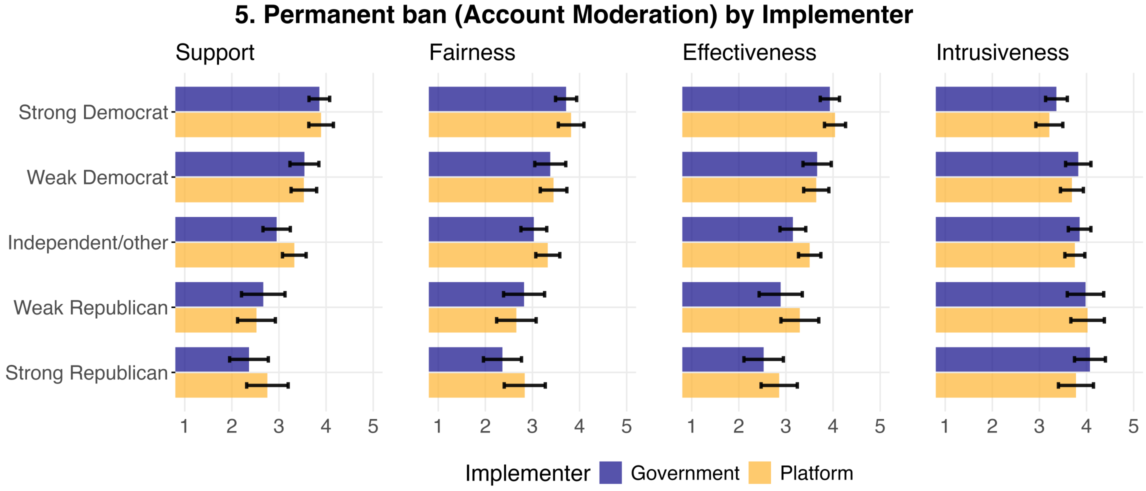


Figure H.5: Average Likert ratings for the **account removal** intervention: permanently banning users who post misinformation a certain number of times.

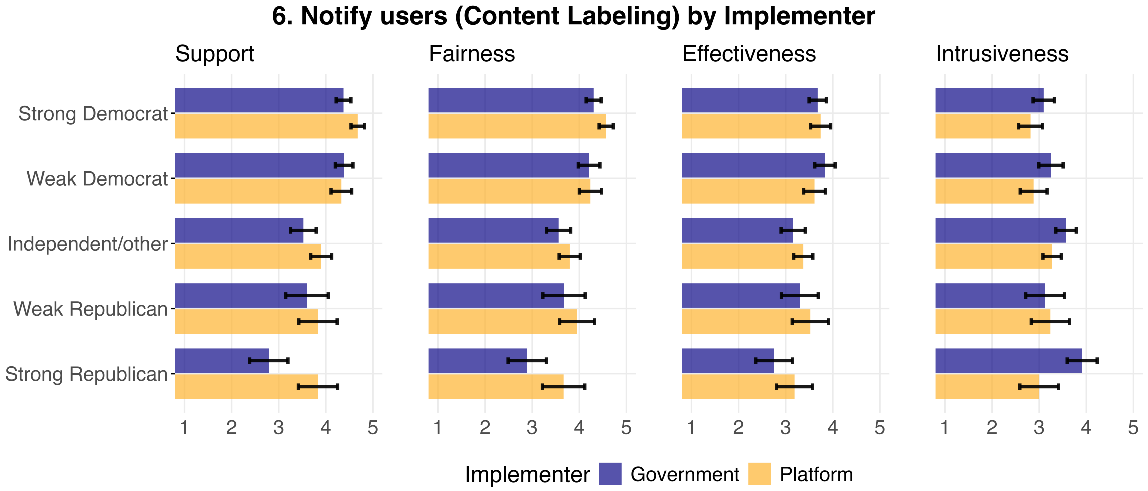


Figure H.6: Average Likert ratings for the **misinformation disclosure** intervention: notifying users if they posted content verified to contain misinformation.

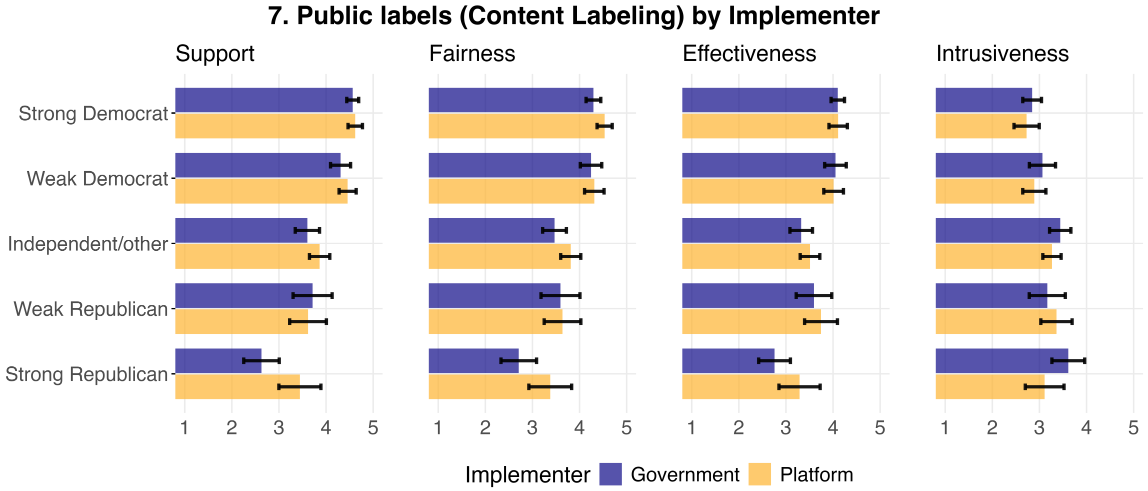


Figure H.7: Average Likert ratings for the **fact-check labelling** intervention: publicly labelling posts verified to contain misinformation with information about and from verified sources.

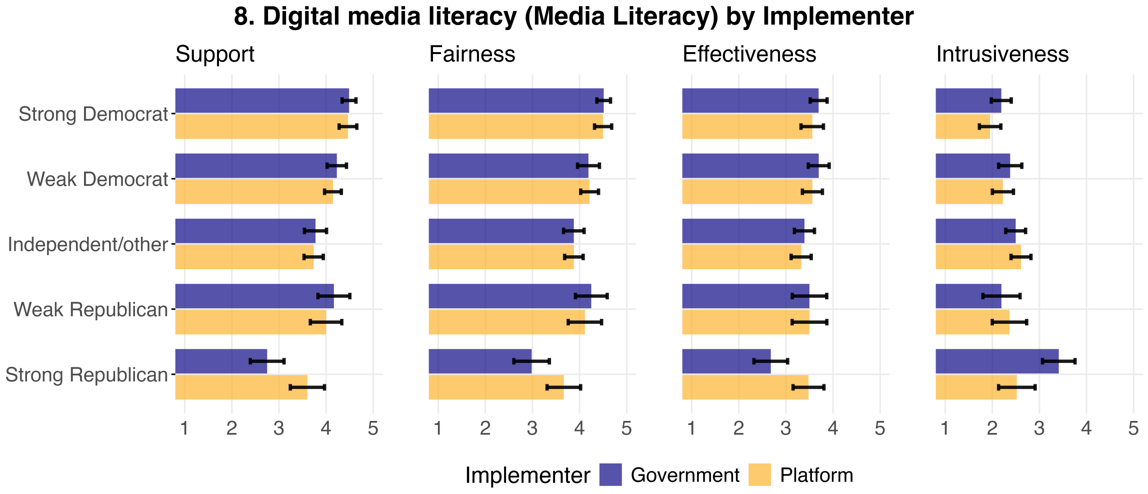


Figure H.8: Average Likert ratings for the **media literacy** intervention: investing in digital media literacy and promoting educational content about detecting misinformation on and offline.

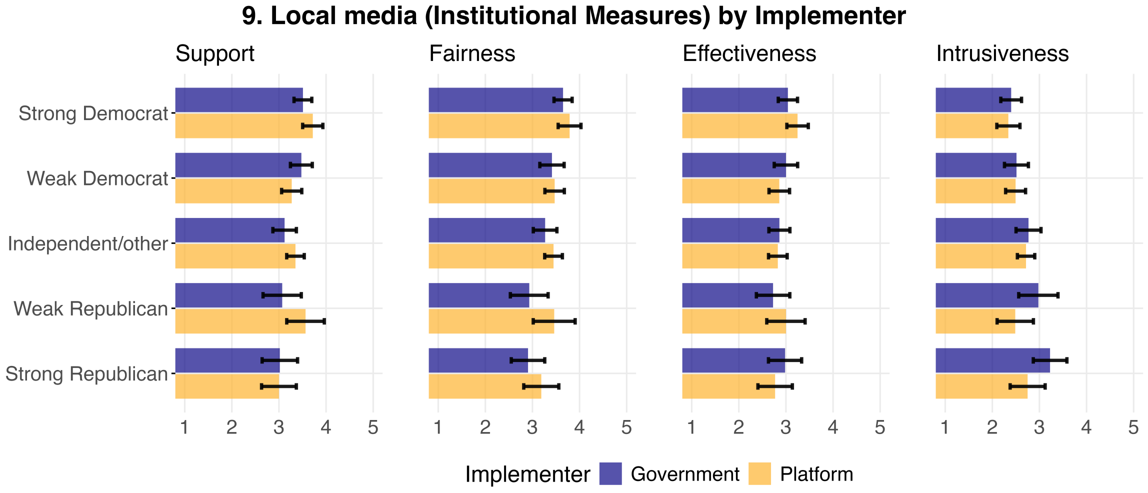


Figure H.9: Average Likert ratings for the **media support** intervention: promoting and investing in local media, which is thought to be most in tune with local norms, culture, and context.

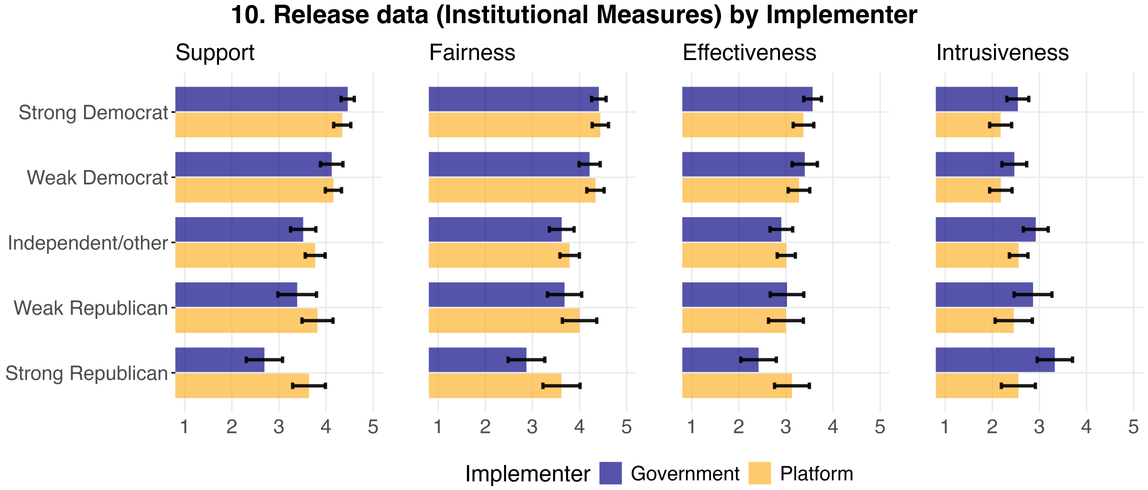


Figure H.10: Average Likert ratings for the **data sharing** intervention: regularly releasing data and/or internal research reports about misinformation to the public and outside researchers.

H.2 One-way ANOVA Statistical Tests

Table H.3 shows the results (F-statistics and degrees of freedom) from the one-way ANOVA tests comparing support, perceived fairness, perceived effectiveness and perceived intrusiveness across categories for each demographic variable. Categories were collapsed such that there were at least 50 participants in each category, except the "Other" gender category which only had 25 participants. This test is used to determine whether there were differences in support levels or perceptions among demographic groups.

Table H.3: F-statistics from one-way ANOVA tests comparing average ratings for support, perceived fairness, perceived effectiveness, and perceived intrusiveness across categories for each demographic variable. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

	Categories	Support	Fairness	Effectiveness	Intrusiveness
Age	18–34	$F(4, 1005) =$ 1.006	$F(4, 1005) =$ 1.268	$F(4, 1005) =$ 1.686	$F(4, 1005) =$ 2.707^{*}
	35–44				
	45–54				
	55–64				
	65+				
Gender	Women	$F(2, 1007) =$ 8.980^{***}	$F(2, 1007) =$ 4.992^{**}	$F(2, 1007) =$ 10.282^{***}	$F(2, 1007) =$ 2.760
	Men				
	Other				
Education	High school or less	$F(4, 1005) =$ 3.027^{*}	$F(4, 1005) =$ 2.094	$F(4, 1005) =$ 2.029	$F(4, 1005) = 0.205$ 0.205
	Some college				
	Associate's degree				
	Bachelor's degree				
	Master's degree or higher				
Income	Less than $20,000	$F(6, 1003) =$ 1.400	$F(6, 1003) =$ 1.636	$F(6, 1003) =$ 0.806	$F(6, 1003) =$ 1.132
	$20,000–$39,999				
	$40,000–$59,999				
	$60,000–$79,999				
	$80,000–$99,999				
	$100,000–$149,999				
	Over $150,000				
Party	Strong Democrat	$F(4, 1005) =$ 61.972^{***}	$F(4, 1005) =$ 51.508^{***}	$F(4, 1005) =$ 34.935^{***}	$F(4, 1005) =$ 21.284^{***}
	Weak Democrat				
	Independent/other				
	Weak Republican				
	Strong Republican				
Ideology	Very liberal	$F(4, 1005) =$ 71.889^{***}	$F(4, 1005) =$ 58.054^{***}	$F(4, 1005) =$ 36.968^{***}	$F(4, 1005) =$ 21.980^{***}
	Liberal				
	Moderate				
	Conservative				
	Very conservative				
Exposure	Never	$F(4, 1005) =$ 0.177	$F(4, 1005) =$ 0.360	$F(4, 1005) =$ 0.469	$F(4, 1005) =$ 0.479
	Rarely				
	Sometimes				
	Often				
	Very Often				

Bibliography

[1] Fraud Detection. URL https://www.qualtrics.com/support/survey-platform/survey-module/survey-checker/fraud-detection/. 4, 5

[2] Auto Bailout Now Backed, Stimulus Divisive. Technical report, Pew Research Center, February 2012. URL https://www.pewresearch.org/politics/2012/02/23/auto-bailout-now-backed-stimulus-divisive/. 5.5

[3] Many believe misinformation is increasing extreme political views and behaviors. Technical report, The AP-NORC Center for Public Affairs Research, October 2022. URL https://apnorc.org/projects/many-believe-misinformation-is-increasing-extreme-political-views-and-behaviors/. 1

[4] Privacy of Employee and Student Social Media Accounts. Technical report, NCSL, August 2022. URL https://www.ncsl.org/technology-and-communication/privacy-of-employee-and-student-social-media-accounts. 7.4.3

[5] Fact Check: Earth is not flat or surrounded by an ice wall. *Reuters*, September 2023. URL https://www.reuters.com/fact-check/earth-is-not-flat-or-surrounded-by-an-ice-wall-2023-09-29/. 1.3.2

[6] Democracy by Design: Social Media's Policy Scores. Technical report, Accountable Tech, 2024. URL https://accountabletech.org/research/democracy-by-design-social-medias-policy-scores/. 6.2.1, 6.4.1, 6.4.2, A.1, A.2

[7] What Is Prompt Engineering? Definition and Examples, October 2024. URL https://www.coursera.org/articles/what-is-prompt-engineering. 2.2.5

[8] Zhila Aghajari, Eric P. S. Baumer, and Dominic DiFranzo. 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*, pages 87:1–87:34. Association for Computing Machinery, April 2023. doi: 10.1145/3579520. URL https://doi.org/10.1145/3579520. 1.4.1, 1.4, 1.4.1, 2.1, 2.2, 7.1.1

[9] Kathryn J. Aikin, Brian G. Southwell, Ryan S. Paquin, Douglas J. Rupert, Amie C. O'Donoghue, Kevin R. Betts, and Philip K. Lee. Correction of misleading information in prescription drug television advertising: The roles of advertisement similarity and time delay. *Research in Social and Administrative Pharmacy*, 13(2):378–388, March 2017. ISSN 1551-7411. doi: 10.1016/j.sapharm.2016.04.004. URL https://www.sciencedirect.com/science/

article/pii/S1551741116300146. 2.3, 6.4.1, A.1

[10] 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 *13th ACM Web Science Conference 2021*, pages 187–195, Virtual Event United Kingdom, June 2021. ACM. ISBN 978-1-4503-8330-1. doi: 10.1145/3447535.3462637. URL https://dl.acm.org/doi/10.1145/3447535.3462637. 7.3.4

[11] Jennifer Allen and David Rand. Combating Misinformation Runs Deeper Than Swatting Away 'Fake News', September 2024. URL https://www.scientificamerican.com/article/combating-misinformation-runs-deeper-than-swatting-away-fake-news/. 5.1

[12] Jennifer Allen, Antonio A. Arechar, Gordon Pennycook, and David G. Rand. Scaling up fact-checking using the wisdom of crowds. *Science Advances*, 7(36):eabf4393, September 2021. doi: 10.1126/sciadv.abf4393. URL https://www.science.org/doi/full/10.1126/sciadv.abf4393. A.4

[13] 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 crowd-sourced 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. URL https://dl.acm.org/doi/10.1145/3491102.3502040. A.4

[14] Bobby Allyn. The Rise of 'Grid Zero': Why more Instagram users are hiding their profile. *NPR*, April 2024. URL https://www.npr.org/2024/04/19/1245316400/grid-zero-instagram-users-hiding-profile-privacy. 3.4.1

[15] Sacha Altay and Fabrizio Gilardi. People are skeptical of headlines labeled as AI-generated, even if true or human-made, because they assume full AI automation. *PNAS Nexus*, 3(10):pgae403, October 2024. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgae403. URL https://doi.org/10.1093/pnasnexus/pgae403. 7.3.5

[16] Sacha Altay, Manon Berriche, Hendrik Heuer, Johan Farkas, and Steven Rathje. A survey of expert views on misinformation: Definitions, determinants, solutions, and future of the field. *Harvard Kennedy School Misinformation Review*, July 2023. doi: 10.37016/mr-2020-119. URL https://misinforeview.hks.harvard.edu/article/a-survey-of-expert-views-on-misinformation-definitions-determinants-solutions-and-future-of-the-field/. 6.4, 6.4.3

[17] Antonio A. Arechar, Jennifer Allen, Adam J. Berinsky, Rocky Cole, Ziv Epstein, Kiran Garimella, Andrew Gully, Jackson G. Lu, Robert M. Ross, Michael N. Stagnaro, Yunhao Zhang, Gordon Pennycook, and David G. Rand. Understanding and combatting misinformation across 16 countries on six continents. *Nature Human Behaviour*, 7(9):1502–1513, September 2023. ISSN 2397-3374. doi: 10.1038/s41562-023-01641-6. URL https://www.nature.com/articles/s41562-023-01641-6. 2.3

[18] Ahmer Arif, John J. Robinson, Stephanie A. Stanek, Elodie S. Fichet, Paul Townsend,

Zena Worku, and Kate Starbird. 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, pages 155–168, New York, NY, USA, February 2017. Association for Computing Machinery. ISBN 978-1-4503-4335-0. doi: 10.1145/2998181.2998294. URL `https://dl.acm.org/doi/10.1145/2998181.2998294`. A.5

[19] K. Peren Arin, Deni Mazrekaj, and Marcel Thum. Ability of detecting and willingness to share fake news. *Scientific Reports*, 13(1):7298, May 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-34402-6. URL `https://www.nature.com/articles/s41598-023-34402-6`. 3.1

[20] Elliot Aronson, Carrie Fried, and Jeff 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, 7.3.1

[21] Kevin Aslett, Zeve Sanderson, William Godel, Nathaniel Persily, Jonathan Nagler, and Joshua A. Tucker. Online searches to evaluate misinformation can increase its perceived veracity. *Nature*, 625(7995):548–556, January 2024. ISSN 1476-4687. doi: 10.1038/s41586-023-06883-y. URL `https://www.nature.com/articles/s41586-023-06883-y`. 6.4.6, 7.3.4

[22] 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. URL `https://doi.org/10.1177/08944393211012268`. 5.1, 6.4.7, A.7

[23] Brooke Auxier. 54% of Americans say social media companies shouldn't allow any political ads, September 2020. URL `https://www.pewresearch.org/fact-tank/2020/09/24/54-of-americans-say-social-media-companies-shouldnt-allow-any-political-ads/`. 6.4.1

[24] Brooke Auxier. 64% of Americans say social media have a mostly negative effect on the way things are going in the U.S. today, October 2020. URL `https://www.pewresearch.org/short-reads/2020/10/15/64-of-americans-say-social-media-have-a-mostly-negative-effect-on-the-way-things-are-going-in-the-u-s-today/`. 3.3.5

[25] Brooke Auxier and Monica Anderson. Social Media Use in 2021, April 2021. URL `https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/`. 3.2.3

[26] Mihai Avram, Nicholas Micallef, Sameer Patil, and Filippo Menczer. Exposure to social engagement metrics increases vulnerability to misinformation. *Harvard Kennedy School Misinformation Review*, July 2020. doi: 10.37016/mr-2020-033.

URL https://misinforeview.hks.harvard.edu/article/exposure-to-social-engagement-metrics-increases-vulnerability-to-misinformation/. 3.5, 7.4.2, A.5

[27] Sumitra Badrinathan. Educative Interventions to Combat Misinformation: Evidence from a Field Experiment in India. *American Political Science Review*, 115(4):1325–1341, November 2021. ISSN 0003-0554, 1537-5943. doi: 10.1017/S0003055421000459. URL https://www.cambridge.org/core/journals/american-political-science-review/article/educative-interventions-to-combat-misinformation-evidence-from-a-field-experiment-in-india/A522EB5164406DE320647014946D31B3. 2.5, 4.2.1

[28] 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*, 29: 394–416, February 2023. ISSN 1940-1612. doi: 10.1177/19401612231158770. 1.4.1, 2.3, 3, 4.2.2, 5.1, A.5

[29] Bence Bago, David G. Rand, and Gordon Pennycook. Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of Experimental Psychology: General*, 149(8):1608–1613, 2020. ISSN 1939-2222. doi: 10.1037/xge0000729. 2.3, 6.4.1, A.1

[30] Joseph B. Bak-Coleman, Ian Kennedy, Morgan Wack, Andrew Beers, Joseph S. Schafer, Emma S. Spiro, Kate Starbird, and Jevin D. West. Combining interventions to reduce the spread of viral misinformation. *Nature Human Behaviour*, 6(10):1–9, June 2022. ISSN 2397-3374. doi: 10.1038/s41562-022-01388-6. 5.1, A.8

[31] Sebastian Bamberg and Daniel Rölle. Determinants of People's Acceptability of Pricing Measures – Replication and Extension of a Causal Model. In Jens Schade and Bernhard Schlag, editors, *Acceptability of Transport Pricing Strategies*, pages 235–248. Emerald Group Publishing Limited, January 2003. ISBN 978-0-08-044199-3 978-1-78635-950-6. doi: 10.1108/9781786359506-015. URL https://doi.org/10.1108/9781786359506-015. 5.5

[32] Anton Barbashin. Improving the Western Strategy to Combat Kremlin Propaganda and Disinformation. Technical report, Atlantic Council: Eurasia Center, May 2018. URL https://www.atlanticcouncil.org/wp-content/uploads/2018/06/Improving_the_Western_Strategy.pdf. 5.1, A.7

[33] Toby Bargar. The Next Horizon for Digital Advertising Taxes Around the Country, January 2023. URL https://news.bloombergtax.com/tax-insights-and-commentary/the-next-horizon-for-digital-advertising-taxes-around-the-country. 7.4.3

[34] Paul Barrett. Tackling Domestic Disinformation: What the Social Media Companies Need to Do. Technical report, NYU Stern Center for Business and Human Rights, March 2019. URL https://issuu.com/nyusterncenterforbusinessandhumanri/docs/nyu_domestic_disinformation_digital. 5.1, 7.4.3, 7.4.4, 7.4.4, A.1, A.2, A.7

[35] Michael Barthel, Amy Mitchell, and Jesse Holcomb. Many Americans Believe Fake News Is Sowing Confusion, December 2016. URL https://www.pewresearch.org/journalism/2016/12/15/many-americans-believe-fake-news-is-sowing-confusion/. 1, 3.1, 3.3.2, 3.3.5

[36] 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. 2.5, 4.1, 4.2.1

[37] Jon Bateman and Dean Jackson. Countering Disinformation Effectively: An Evidence-Based Policy Guide. Technical report, Carnegie Endowment for International Peace, January 2024. URL https://carnegieendowment.org/research/2024/01/countering-disinformation-effectively-an-evidence-based-policy-guide?lang=en. 6.4.6, 6.4.7, 7.4.2, 7.4.3, 7.6

[38] C. Daniel Batson, Diane Kobrynowicz, Jessica L. Dinnerstein, Hannah C. Kampf, and Angela D. Wilson. In a very different voice: Unmasking moral hypocrisy. *Journal of Personality and Social Psychology*, 72(6):1335–1348, 1997. ISSN 1939-1315. doi: 10.1037/0022-3514.72.6.1335. 3.3.1

[39] Nick Beake. Facebook admits it was used to 'incite offline violence' in Myanmar. *BBC*, November 2018. URL https://www.bbc.com/news/world-asia-46105934. 7.3.3

[40] Matthew C. Benigni, Kenneth Joseph, and Kathleen M. Carley. Online extremism and the communities that sustain it: Detecting the ISIS supporting community on Twitter. *PLOS ONE*, 12(12):e0181405, December 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0181405. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0181405. 7.3.3

[41] Rony Berger, Joy Benatov, Hisham Abu-Raiya, and Carmit T. Tadmor. Reducing prejudice and promoting positive intergroup attitudes among elementary-school children in the context of the Israeli–Palestinian conflict. *Journal of School Psychology*, 57:53–72, August 2016. ISSN 0022-4405. doi: 10.1016/j.jsp.2016.04.003. URL https://www.sciencedirect.com/science/article/pii/S0022440516300097. 3

[42] Magnus Bergquist, Andreas Nilsson, Niklas Harring, and Sverker C. Jagers. Meta-analyses of fifteen determinants of public opinion about climate change taxes and laws. *Nature Climate Change*, 12(3):235–240, March 2022. ISSN 1758-6798. doi: 10.1038/s41558-022-01297-6. URL https://www.nature.com/articles/s41558-022-01297-6. 5, 5.3.1, 5.5

[43] Samuel Bestvater. How U.S. Adults Use TikTok, February 2024. URL https://www.pewresearch.org/internet/2024/02/22/how-u-s-adults-use-tiktok/. 3.4.1

[44] 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. Emer-

ald Publishing Limited, January 2017. ISBN 978-1-78714-486-6 978-1-78714-485-9. https://doi.org/10.1108/S2398-601820180000002007. 1, A.7

[45] Robert A. Blair, Jessica Gottlieb, Brendan Nyhan, Laura Paler, Pablo Argote, and Charlene J. Stainfield. Interventions to counter misinformation: Lessons from the Global North and applications to the Global South. Technical report, USAID, July 2023. URL https://democratic-erosion.org/wp-content/uploads/2024/09/INTERVENTIONS-TO-COUNTER-MISINFORMATION-LESSONS-FROM-THE-GLOBAL-NORTH-AND-APPLICATIONS-TO-THE-GLOBAL-SOUTH.pdf. 6.4.1, 6.4.4, 6.4.6

[46] Robert A. Blair, Jessica Gottlieb, Brendan Nyhan, Laura Paler, Pablo Argote, and Charlene J. Stainfield. Interventions to counter misinformation: Lessons from the Global North and applications to the Global South. *Current Opinion in Psychology*, 55:101732, February 2024. ISSN 2352-250X. doi: 10.1016/j.copsyc.2023.101732. URL https://www.sciencedirect.com/science/article/abs/pii/S2352250X2300177X. 1.4.1, 1.4, 1.4.1, 1.4.3, 2.1, 2.5, 4.1, 6.2.4, 6.4, 6.4.1, 6.4.4, 6.4.5, 6.4.6, 6.4.7, 7.1.1

[47] Leticia Bode and Emily K. Vraga. In Related News, That Was Wrong: The Correction of Misinformation Through Related Stories Functionality in Social Media. *Journal of Communication*, 65(4):619–638, August 2015. ISSN 0021-9916. doi: 10.1111/jcom.12166. 2.3, 6.4.1, A.1

[48] 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. 1.4.1, 3, 3.1, 4.2.2, A.2, A.5

[49] Andrew Booth, Diana Papaioannou, and Anthea Sutton. *Systematic Approaches to a Successful Literature Review*. Sage Publications, 1st edition, January 2012. ISBN 978-1-4739-1245-8. 2.2.4

[50] 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. URL https://osf.io/a7huq. A.2

[51] Colin Bos, Ivo Van Der Lans, Frank Van Rijnsoever, and Hans Van Trijp. Consumer Acceptance of Population-Level Intervention Strategies for Healthy Food Choices: The Role of Perceived Effectiveness and Perceived Fairness. *Nutrients*, 7(9):7842–7862, September 2015. ISSN 2072-6643. doi: 10.3390/nu7095370. URL https://www.mdpi.com/2072-6643/7/9/5370. 5, 5.5

[52] Samantha Bradshaw and Lisa-Maria Neudert. The Road Ahead: Mapping Civil Society Responses to Disinformation. Technical report, National Endowment for Democracy, January 2021. URL https://www.ned.org/wp-content/uploads/2021/01/The-Road-Ahead-Mapping-Civil-Society-Responses-to-Disinformation-Bradshaw-Neudert-Jan-2021-2.pdf. 5.1, 6.4.7, 7.4.4,

7.4.4, A.7

[53] Petter Bae Brandtzæg. Social Networking Sites: Their Users and Social Implications — A Longitudinal Study. *Journal of Computer-Mediated Communication*, 17(4):467–488, July 2012. ISSN 1083-6101. doi: 10.1111/j.1083-6101.2012.01580.x. URL `https://doi.org/10.1111/j.1083-6101.2012.01580.x`. 7.4.1

[54] Nadia M. Brashier. Fighting misinformation among the most vulnerable users. *Current Opinion in Psychology*, 57:101813, June 2024. ISSN 2352-250X. doi: 10.1016/j.copsyc.2024.101813. URL `https://www.sciencedirect.com/science/article/pii/S2352250X24000265`. 3.5

[55] Nadia M. Brashier and Daniel L. Schacter. Aging in an Era of Fake News. *Current Directions in Psychological Science*, 29(3):316–323, June 2020. ISSN 0963-7214. doi: 10.1177/0963721420915872. URL `https://doi.org/10.1177/0963721420915872`. 3.3.1, 3.5

[56] Nadia M. Brashier, Gordon Pennycook, Adam J. Berinsky, and David G. Rand. Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences*, 118(5):e2020043118, February 2021. doi: 10.1073/pnas.2020043118. URL `https://www.pnas.org/doi/full/10.1073/pnas.2020043118`. 6.4.6

[57] Aengus Bridgman, Eric Merkley, Peter John Loewen, Taylor Owen, Derek Ruths, Lisa Teichmann, and Oleg Zhilin. The causes and consequences of COVID-19 misperceptions: Understanding the role of news and social media. *Harvard Kennedy School Misinformation Review*, June 2020. doi: 10.37016/mr-2020-028. URL `https://misinforeview.hks.harvard.edu/?p=1832`. 3.3.1

[58] Hendrik Bruns, François J. Dessart, Michał Krawczyk, Stephan Lewandowsky, Myrto Pantazi, Gordon Pennycook, Philipp Schmid, and Laura Smillie. 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, September 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-71599-6. URL `https://www.nature.com/articles/s41598-024-71599-6`. 2.5

[59] Nils Brunsson. *The Organization of Hypocrisy: Talk, Decisions and Actions in Organizations*. Wiley, New York, NY, 1989. 3.1

[60] Hanne Bruun. The delay economy of "continuity" and the emerging impatience culture of the digital era. *Nordic Journal of Media Studies*, 1(1):85–101, June 2019. doi: 10.2478/njms-2019-0006. URL `https://sciendo.com/article/10.2478/njms-2019-0006`. 5.5

[61] Ceren Budak, Brendan Nyhan, David M. Rothschild, Emily Thorson, and Duncan J. Watts. Misunderstanding the harms of online misinformation. *Nature*, 630(8015):45–53, June 2024. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-024-07417-w. URL `https://www.nature.com/articles/s41586-024-07417-w`. 5.1

[62] Cody Buntain, Richard Bonneau, Jonathan Nagler, and Joshua A. Tucker. YouTube Recommendations and Effects on Sharing Across Online Social Platforms. In *Proceedings of the ACM on Human-Computer Interaction*, volume 5 of *CSCW*, pages 11:1–11:26,

April 2021. doi: 10.1145/3449085. URL https://dl.acm.org/doi/10.1145/3449085. 6.4.2

[63] 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. 5, 5.1

[64] John M. Carey, Victoria Chi, D. J. Flynn, Brendan Nyhan, and Thomas Zeitzoff. The effects of corrective information about disease epidemics and outbreaks: Evidence from Zika and yellow fever in Brazil. *Science Advances*, 6(5):eaaw7449, January 2020. doi: 10.1126/sciadv.aaw7449. URL https://www.science.org/doi/10.1126/sciadv.aaw7449. 3.3.1

[65] 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*, pages 1–10. Springer, New York, NY, 2017. ISBN 978-1-4614-7163-9. doi: 10.1007/978-1-4614-7163-9_309-1. URL https://doi.org/10.1007/978-1-4614-7163-9_309-1. 2.3

[66] 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, 7

[67] Sean Carlin. Coronavirus Fears Haven't Sunk Sales of Corona Beer in U.S., March 2020. URL https://www.factcheck.org/2020/03/coronavirus-fears-havent-sunk-sales-of-corona-beer-in-u-s/. 1.3.2

[68] 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, A.2

[69] Xinran Chen, Sei-Ching Joanna Sin, Yin-Leng Theng, and Chei Sian Lee. Why Students Share Misinformation on Social Media: Motivation, Gender, and Study-level Differences. *The Journal of Academic Librarianship*, 41(5):583–592, September 2015. ISSN 0099-1333. doi: 10.1016/j.acalib.2015.07.003. URL https://www.sciencedirect.com/science/article/pii/S0099133315001494. 3.1

[70] Lesley Chiou and Catherine Tucker. Fake News and Advertising on Social Media: A Study of the Anti-Vaccination Movement, November 2018. URL https://www.nber.org/papers/w25223. A.1

[71] Yuwei Chuai, Haoye Tian, Nicolas Pröllochs, and Gabriele Lenzini. Did the Roll-Out of Community Notes Reduce Engagement With Misinformation on X/Twitter? *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW2):428:1–428:52, November 2024. doi: 10.1145/3686967. URL https://dl.acm.org/doi/10.1145/3686967. 2.5, 5.6.1, 6.4.4

[72] Giovanni Luca Ciampaglia. The Digital Misinformation Pipeline. In Olga Zlatkin-Troitschanskaia, Gabriel Wittum, and Andreas Dengel, editors, *Positive Learning in the*

Age of Information: A Blessing or a Curse?, pages 413–421. Springer Fachmedien, Wiesbaden, 2018. ISBN 978-3-658-19567-0. doi: 10.1007/978-3-658-19567-0_25. URL https://doi.org/10.1007/978-3-658-19567-0_25. 1.3.3, 1.3.3, 1.3.3, 1.3.3, 7.1.1

[73] 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. 5.2

[74] Mitchell Clark. Facebook wants to make sure you've read the article you're about to share, May 2021. URL https://www.theverge.com/2021/5/10/22429174/facebook-article-popup-read-misinformation. 3.5, 5.5

[75] Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, April 1960. ISSN 0013-1644. doi: 10.1177/001316446002000104. URL https://doi.org/10.1177/001316446002000104. 2.2.5

[76] Jonas Colliander. "This is fake news": Investigating the role of conformity to other users' views when commenting on and spreading disinformation in social media. *Computers in Human Behavior*, 97:202–215, August 2019. ISSN 07475632. doi: 10.1016/j.chb.2019.03.032. URL https://linkinghub.elsevier.com/retrieve/pii/S074756321930130X. 3.5, 7.4.2

[77] Lorne Cook. EU to limit political ads, ban use of certain personal info, November 2021. URL https://www.pbs.org/newshour/world/eu-to-limit-political-ads-ban-use-of-certain-personal-info. 7.4.3

[78] Thomas H. Costello, Gordon Pennycook, and David G. Rand. Durably reducing conspiracy beliefs through dialogues with AI. *Science*, 385(6714):eadq1814, September 2024. doi: 10.1126/science.adq1814. URL https://www.science.org/doi/10.1126/science.adq1814. 6.4.8, 7.3.5, A.8

[79] 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, 1.4.1, 1.4, 1.4.1, 2.1, 2.4.1, 2.4.3, 2.5, 2.6.1, 3.1, 4.1, 5.1, 5.2.1, 5.1, 7.1.1, A.1, A.3, A.4

[80] Cathy Cranston. Guides: Evaluating Online Information: Home. URL https://www.lib.uiowa.edu/. 1.3.1

[81] Caroline Crystal. Facebook, Telegram, and the Ongoing Struggle Against Online Hate Speech. Technical report, Carnegie Endowment for International Peace, September 2023. URL https://carnegieendowment.org/research/2023/09/facebook-telegram-and-the-ongoing-struggle-against-online-hate-speech?lang=en. 7.3.3, 7.4.2

[82] Stéphane Côté, Julian House, and Robb Willer. High economic inequality leads higher-income individuals to be less generous. *Proceedings of the National Academy of Sciences*,

112(52):15838–15843, December 2015. doi: 10.1073/pnas.1511536112. URL https://www.pnas.org/doi/10.1073/pnas.1511536112. 3.3.5

[83] Brittany I. Davidson, Darja Wischerath, Daniel Racek, Douglas A. Parry, Emily Godwin, Joanne Hinds, Dirk van der Linden, Jonathan F. Roscoe, Laura Ayravainen, and Alicia G. Cork. Platform-controlled social media APIs threaten open science. *Nature Human Behaviour*, 7(12):2054–2057, December 2023. ISSN 2397-3374. doi: 10.1038/s41562-023-01750-2. URL https://www.nature.com/articles/s41562-023-01750-2. 7.4.4

[84] David De Coninck, Thomas Frissen, Koen Matthijs, Leen d'Haenens, Grégoire Lits, Olivier Champagne-Poirier, Marie-Eve Carignan, Marc D. David, Nathalie Pignard-Cheynel, Sébastien Salerno, and Melissa Généreux. Beliefs in Conspiracy Theories and Misinformation About COVID-19: Comparative Perspectives on the Role of Anxiety, Depression and Exposure to and Trust in Information Sources. *Frontiers in Psychology*, 12:646394, April 2021. ISSN 1664-1078. doi: 10.3389/fpsyg.2021.646394. URL https://www.frontiersin.org/articles/10.3389/fpsyg.2021.646394/full. 5.1

[85] Jenny de Fine Licht. Policy Area as a Potential Moderator of Transparency Effects: An Experiment. *Public Administration Review*, 74(3):361–371, 2014. ISSN 1540-6210. doi: 10.1111/puar.12194. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/puar.12194. 5.5

[86] Ratan Dey, Zubin Jelveh, and Keith Ross. Facebook users have become much more private: A large-scale study. In *2012 IEEE International Conference on Pervasive Computing and Communications Workshops*, pages 346–352, March 2012. doi: 10.1109/PerComW.2012.6197508. URL https://ieeexplore.ieee.org/abstract/document/6197508. 3.4.1

[87] Stephanie Diepeveen, Tom Ling, Marc Suhrcke, Martin Roland, and Theresa M. Marteau. Public acceptability of government intervention to change health-related behaviours: a systematic review and narrative synthesis. *BMC Public Health*, 13(1):756, August 2013. ISSN 1471-2458. doi: 10.1186/1471-2458-13-756. URL https://doi.org/10.1186/1471-2458-13-756. 5, 5.5, 5.6.1, 6.2.3, 6.2.3, 7.3.2

[88] Stacy Jo Dixon. U.S. social media user account privacy 2018, 2018. URL https://www.statista.com/statistics/934874/users-have-private-social-media-account-usa/. 3.4.1

[89] 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/. 5, 5.1

[90] Chiara Patricia Drolsbach, Kirill Solovev, and Nicolas Pröllochs. Community notes increase trust in fact-checking on social media. *PNAS Nexus*, 3(7):pgae217, July 2024. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgae217. URL https://doi.org/10.1093/pnasnexus/pgae217. 5.6.1, 6.4.4

[91] James N. Druckman, Katherine Ognyanova, Matthew A. Baum, David Lazer, Roy H. Perlis, John Della Volpe, Mauricio Santillana, Hanyu Chwe, Alexi Quintana, and Matthew Simonson. The role of race, religion, and partisanship in misperceptions about COVID-19. *Group Processes & Intergroup Relations*, 24(4):638–657, June 2021. ISSN 1368-4302. doi: 10.1177/1368430220985912. URL https://doi.org/10.1177/1368430220985912. 3.3.1

[92] Susan E. Dudley and Jerry Brito, editors. *Regulation: A Primer*. Mercatus Center at George Mason University, Arlington, Va, 2nd ed edition, 2012. ISBN 978-0-9836077-3-1 978-0-9836077-4-8. 6.2.4

[93] Francesca D'Errico, Paolo Giovanni Cicirelli, Giuseppe Corbelli, and Marinella Paciello. Addressing racial misinformation at school: a psycho-social intervention aimed at reducing ethnic moral disengagement in adolescents. *Social Psychology of Education*, 27 (3):611–630, June 2024. ISSN 1573-1928. doi: 10.1007/s11218-023-09777-z. URL https://doi.org/10.1007/s11218-023-09777-z. 3.3.1, 3.5

[94] Ullrich Ecker, Jon Roozenbeek, Sander van der Linden, Li Qian Tay, John Cook, Naomi Oreskes, and Stephan Lewandowsky. Misinformation remains a threat to democracy. *Nature*, 630:29–32, June 2024. doi: 10.1038/d41586-024-01587-3. URL https://www.nature.com/articles/d41586-024-01587-3. 5.1

[95] Ullrich K. H. Ecker and Luke M. Antonio. Can you believe it? An investigation into the impact of retraction source credibility on the continued influence effect. *Memory & Cognition*, 49(4):631–644, May 2021. ISSN 1532-5946. doi: 10.3758/s13421-020-01129-y. URL https://doi.org/10.3758/s13421-020-01129-y. 2.3, A.5

[96] 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

[97] Ullrich K. H. Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K. Fazio, Nadia Brashier, Panayiota Kendeou, Emily K. Vraga, and Michelle A. Amazeen. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1):13–29, January 2022. ISSN 2731-0574. doi: 10.1038/s44159-021-00006-y. URL https://www.nature.com/articles/s44159-021-00006-y. 3.1

[98] Ullrich K. H. Ecker, Jasmyne A. Sanderson, Paul McIlhiney, Jessica J. Rowsell, Hayley L. Quekett, Gordon DA Brown, and Stephan Lewandowsky. Combining refutations and social norms increases belief change. *Quarterly Journal of Experimental Psychology (2006)*, 76(6):1275–1297, June 2023. ISSN 1747-0226. doi: 10.1177/17470218221111750. URL https://journals.sagepub.com/doi/10.1177/17470218221111750. 2.3, 6.4.5, A.5

[99] Vittoria Elliott. In Bulgaria, Russian Trolls Are Winning the Information War. *Wired*, March 2023. ISSN 1059-1028. URL https://www.wired.com/story/in-bulgaria-russian-trolls-are-winning-the-information-war/.

7.3.4

[100] Adam M. Enders, Joseph E. Uscinski, Michelle I. Seelig, Casey A. Klofstad, Stefan Wuchty, John R. Funchion, Manohar N. Murthi, Kamal Premaratne, and Justin Stoler. The Relationship Between Social Media Use and Beliefs in Conspiracy Theories and Misinformation. *Political Behavior*, 45(2):781–804, June 2023. ISSN 0190-9320, 1573-6687. doi: 10.1007/s11109-021-09734-6. URL `https://link.springer.com/10.1007/s11109-021-09734-6`. 5.1

[101] Ziv Epstein, Adam J. Berinsky, Rocky Cole, Andrew Gully, Gordon Pennycook, and David G. Rand. Developing an accuracy-prompt toolkit to reduce COVID-19 misinformation online. *Harvard Kennedy School Misinformation Review*, May 2021. doi: 10.37016/mr-2020-71. URL `https://misinforeview.hks.harvard.edu/article/developing-an-accuracy-prompt-toolkit-to-reduce-covid-19-misinformation-online/`. 2.3, A.1

[102] Ziv Epstein, Mengying Cathy Fang, Antonio Alonso Arechar, and David Rand. What label should be applied to content produced by generative AI?, July 2023. URL `https://osf.io/v4mfz_v1`. 7.3.5

[103] 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, 7.3.1

[104] Daniel Funke and Daniela Flamini. A guide to anti-misinformation actions around the world. URL `https://www.poynter.org/ifcn/anti-misinformation-actions/`. 7.4.3

[105] Mingkun Gao, Ziang Xiao, Karrie Karahalios, and Wai-Tat Fu. 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*, pages 55:1–55:16, November 2018. doi: 10.1145/3274324. URL `https://dl.acm.org/doi/10.1145/3274324`. 2.3, A.4

[106] Naman Garg and Monika Yadav. Learning to Resist Misinformation: A Field Experiment in India, 2022. URL `https://www.aeaweb.org/doi/10.1257/rct.7923`. 2.3

[107] R. Kelly Garrett and Robert M. Bond. Conservatives' susceptibility to political misperceptions. *Science Advances*, 7(23):eabf1234, June 2021. doi: 10.1126/sciadv.abf1234. URL `https://www.science.org/doi/full/10.1126/sciadv.abf1234`. 3.3.1

[108] Andrew Gelman, Jennifer Hill, and Masanao Yajima. Why We (Usually) Don't Have to Worry About Multiple Comparisons. *Journal of Research on Educational Effectiveness*, 5(2):189–211, April 2012. ISSN 1934-5747. doi: 10.1080/19345747.2011.618213. URL `https://doi.org/10.1080/19345747.2011.618213`. 3.3.2

[109] Tarleton Gillespie. Do Not Recommend? Reduction as a Form of Content Moderation. *Social Media + Society*, 8(3):20563051221117552, July 2022. ISSN 2056-3051. doi: 10.1177/20563051221117552. URL `https://doi.org/10.1177/20563051221117552`. 5.1, A.1, A.2

[110] Henner Gimpel, Sebastian Heger, Christian Olenberger, and Lena Utz. The Effectiveness of Social Norms in Fighting Fake News on Social Media. *Journal of Management Information Systems*, 38(1):196–221, January 2021. ISSN 0742-1222. doi: 10.1080/07421222.2021.1870389. 3.5, 6.4.5, A.5

[111] Priscilla Glasow. Fundamentals of Survey Research Methodology. Technical Report MP 05W0000077, MITRE, McLean, Virginia, April 2005. URL https://www.uky.edu/~kdbrad2/EPE619/Handouts/SurveyResearchReading.pdf. 7, 7.5

[112] Burt Glass. Survey: Across parties, Americans accept removal of false health info by social media companies, survey says, February 2025. URL https://sites.bu.edu/crc/2025/02/17/survey-across-parties-americans-accept-removal-of-false-health-info-by-social-media-companies-survey-says/. 6.2.4

[113] Laura K Globig, Nora Holtz, and Tali Sharot. Changing the incentive structure of social media platforms to halt the spread of misinformation. *eLife*, 12:e85767, June 2023. ISSN 2050-084X. doi: 10.7554/eLife.85767. URL https://doi.org/10.7554/eLife.85767. 7.3.2, 7.4.2

[114] Sandra González-Bailón, David Lazer, Pablo Barberá, Meiqing Zhang, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Deen Freelon, Matthew Gentzkow, Andrew M. Guess, Shanto Iyengar, Young Mie Kim, Neil Malhotra, Devra Moehler, Brendan Nyhan, Jennifer Pan, Carlos Velasco Rivera, Jaime Settle, Emily Thorson, Rebekah Tromble, Arjun Wilkins, Magdalena Wojcieszak, Chad Kiewiet de Jonge, Annie Franco, Winter Mason, Natalie Jomini Stroud, and Joshua A. Tucker. Asymmetric ideological segregation in exposure to political news on Facebook. *Science*, 381(6656):392–398, July 2023. doi: 10.1126/science.ade7138. URL https://www.science.org/doi/full/10.1126/science.ade7138. 3.3.1

[115] Sandra González-Bailón, David Lazer, Pablo Barberá, William Godel, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Deen Freelon, Matthew Gentzkow, Andrew M. Guess, Shanto Iyengar, Young Mie Kim, Neil Malhotra, Devra Moehler, Brendan Nyhan, Jennifer Pan, Carlos Velasco Rivera, Jaime Settle, Emily Thorson, Rebekah Tromble, Arjun Wilkins, Magdalena Wojcieszak, Chad Kiewiet de Jonge, Annie Franco, Winter Mason, Natalie Jomini Stroud, and Joshua A. Tucker. The Diffusion and Reach of (Mis)Information on Facebook During the U.S. 2020 Election. *Sociological Science*, 11:1124–1146, December 2024. ISSN 2330-6696. doi: 10.15195/v11.a41. URL https://sociologicalscience.com/articles-v11-41-1124/. 6.2.4

[116] 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. URL https://doi.org/10.1177/2053951719897945. A.2

[117] Yasmin Green, Andrew Gully, Yoel Roth, Abhishek Roy, Joshua A. Tucker, and Alicia Wanless. Evidence-Based Misinformation Interventions: Challenges and Opportunities for Measurement and Collaboration. Technical re-

port, Carnegie Endowment for International Peace, December 2022. URL https://carnegieendowment.org/2023/01/09/evidence-based-misinformation-interventions-challenges-and-opportunities-for-measurement-and-collaboration-pub-88661. 6.1, 6.2.3, 6.4.5, 7.1.1

[118] Sonja Grelle and Wilhelm Hofmann. When and Why Do People Accept Public-Policy Interventions? An Integrative Public-Policy-Acceptance Framework. *Perspectives on Psychological Science*, 19(1):258–279, January 2024. ISSN 1745-6916. doi: 10.1177/17456916231180580. URL https://doi.org/10.1177/17456916231180580. 5, 5.3.1, 5.5, 5.6.1, 6.2.2, 6.2.3, 6.2.3, 6.2.4, 6.4.8, 7, 7.4.3, 7.5

[119] Eric Griffith. Social Privacy Is on the Rise: Almost Half of Social Media Accounts Are Kept Private, June 2020. URL https://www.pcmag.com/news/social-privacy-is-on-the-rise-almost-half-of-social-media-accounts-are. 3.4.1

[120] Andrew Guess, Brendan Nyhan, and Jason Reifler. Selective Exposure to Misinformation: Evidence from the consumption of fake news during the 2016 U.S. presidential campaign. Technical report, 2018. URL https://about.fb.com/wp-content/uploads/2018/01/fake-news-2016.pdf. 3.3.1

[121] 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. 2.5, 4.1, 4.2.1, 5.1, 6.2.4, 6.4.6, 7.3.4, A.6

[122] 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. 3.3.1, 5.2.1

[123] Désirée Hagmann, Michael Siegrist, and Christina Hartmann. Taxes, labels, or nudges? Public acceptance of various interventions designed to reduce sugar intake. *Food Policy*, 79:156–165, August 2018. ISSN 0306-9192. doi: 10.1016/j.foodpol.2018.06.008. URL https://www.sciencedirect.com/science/article/pii/S0306919217310096. 5.5, 6.2.3, 7.3.2

[124] Stuart Hall. Encoding and decoding in the television discourse. Technical report, Birmingham: Centre for Contemporary Cultural Studies, 1973. 1.3.2

[125] David J. Hauser and Norbert Schwarz. Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, 48(1):400–407, March 2016. ISSN 1554-3528. doi: 10.3758/s13428-015-0578-z. URL https://doi.org/10.3758/s13428-015-0578-z. 3.2.3

[126] David J. Hauser, Aaron J. Moss, Cheskie Rosenzweig, Shalom N. Jaffe, Jonathan Robinson, and Leib Litman. Evaluating CloudResearch's Approved Group as a solution for problematic data quality on MTurk. *Behavior Research Methods*, 55(8):3953–3964, De-

cember 2023. ISSN 1554-3528. doi: 10.3758/s13428-022-01999-x. URL https://doi.org/10.3758/s13428-022-01999-x. 3.2.3

[127] Danny Hayes and Jennifer L. Lawless. The Decline of Local News and Its Effects: New Evidence from Longitudinal Data. *The Journal of Politics*, 80(1):332–336, January 2018. ISSN 0022-3816. doi: 10.1086/694105. URL https://www.journals.uchicago.edu/doi/full/10.1086/694105. 6.4.7

[128] Bing He, Yibo Hu, Yeon-Chang Lee, Soyoung Oh, Gaurav Verma, and Srijan Kumar. A Survey on the Role of Crowds in Combating Online Misinformation: Annotators, Evaluators, and Creators. *ACM Trans. Knowl. Discov. Data*, 19(1):10:1–10:30, November 2024. ISSN 1556-4681. doi: 10.1145/3694980. URL https://dl.acm.org/doi/10.1145/3694980. 7.4.1

[129] Moreen Heine, Susanna Kuper, and Thomas Neururer. Which Platform to Use?: Social Media Platform Types and their Suitability for Sound Decision Making by Voluntary Helpers. In *Proceedings of the 11th International Conference on Theory and Practice of Electronic Governance*, pages 395–402, Galway Ireland, April 2018. ACM. ISBN 978-1-4503-5421-9. doi: 10.1145/3209415.3209485. URL https://dl.acm.org/doi/10.1145/3209415.3209485. 3.4.1, 3.4.1

[130] Todd C. Helmus and Bilva Chandra. Generative Artificial Intelligence Threats to Information Integrity and Potential Policy Responses. Technical report, RAND Corporation, April 2024. URL https://www.rand.org/pubs/perspectives/PEA3089-1.html. 7.4.3, 7.4.3, 7.4.4, A.8

[131] 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. 1.4.1, 2.5, 4.1, 5.1, 5.2.1, 5.1, A.1

[132] Alex Hern. Facebook lifts ban on posts claiming Covid-19 was man-made. *The Guardian*, May 2021. ISSN 0261-3077. URL https://www.theguardian.com/technology/2021/may/27/facebook-lifts-ban-on-posts-claiming-covid-19-was-man-made. 7.1.1

[133] Jeff Horwitz. Facebook Has Made Lots of New Rules This Year. It Doesn't Always Enforce Them. *WSJ*, October 2020. URL https://www.wsj.com/tech/facebook-has-made-lots-of-new-rules-this-year-it-doesnt-always-enforce-them-11602775676. 7.3.3, 7.4.2

[134] David M. Howcroft and Verena Rieser. What happens if you treat ordinal ratings as interval data? Human evaluations in NLP are even more under-powered than you think. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8932–8939, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.703. URL https://aclanthology.org/2021.emnlp-main.703. 3.3.3

[135] Guanxiong Huang, Wufan Jia, and Wenting Yu. Media Literacy Interventions Improve Resilience to Misinformation: A Meta-Analytic Investigation of Overall Effect and Moder-

ating Factors. *Communication Research*, page 00936502241288103, October 2024. ISSN 0093-6502. doi: 10.1177/00936502241288103. URL https://doi.org/10.1177/00936502241288103. 6.2.4

[136] 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. 5, 5.2, 5.3.1, 5.5

[137] Søren Højsgaard, Ulrich Halekoh, and Jun Yan. The R Package geepack for Generalized Estimating Equations. *Journal of Statistical Software*, 15:1–11, 2006. ISSN 1548-7660. doi: 10.18637/jss.v015.i02. URL https://doi.org/10.18637/jss.v015.i02. D

[138] Jane Im, Sonali Tandon, Eshwar Chandrasekharan, Taylor Denby, and Eric Gilbert. Synthesized Social Signals: Computationally-Derived Social Signals from Account Histories. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI, pages 1–12, New York, NY, USA, April 2020. Association for Computing Machinery. ISBN 978-1-4503-6708-0. doi: 10.1145/3313831.3376383. URL https://dl.acm.org/doi/10.1145/3313831.3376383. 2.5

[139] Paul Jaccard. Étude comparative de la distribution florale dans une portion des Alpes et du Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901. doi: 10.5169/SEALS-266450. URL http://dx.doi.org/10.5169/seals-266450. 2.2.5

[140] Kristin M. Jackson and William M. K. Trochim. Concept Mapping as an Alternative Approach for the Analysis of Open-Ended Survey Responses. *Organizational Research Methods*, 5(4):307–336, October 2002. ISSN 1094-4281. doi: 10.1177/109442802237114. URL https://doi.org/10.1177/109442802237114. 4.3.5

[141] 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. 1.3.3, 1.4.1, 4.1, 4.2.1, 7, A.6

[142] Shagun Jhaver and Amy X. Zhang. 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*, December 2023. ISSN 1461-4448. doi: 10.1177/14614448231217993. URL https://doi.org/10.1177/14614448231217993. 4.2.2, 6.4.2, A.2

[143] 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. URL https://dl.acm.org/doi/10.1145/3534929. 1.4.1, 5.1, 6.2.2, 6.2.3, A.2

[144] Amelia Johns, Francesco Bailo, Emily Booth, and Marian-Andrei Rizoiu. 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*, March 2024. ISSN 1329-878X, 2200-467X. doi: 10.1177/

1329878X241236984. URL https://journals.sagepub.com/doi/10.1177/1329878X241236984. 6.4.3, A.3

[145] Christopher M. Jones, Daniel Diethei, Johannes Schöning, Rehana Shrestha, Tina Jahnel, and Benjamin Schüz. Impact of Social Reference Cues on Misinformation Sharing on Social Media: Series of Experimental Studies. *Journal of Medical Internet Research*, 25(1):e45583, August 2023. doi: 10.2196/45583. URL https://www.jmir.org/2023/1/e45583. 3.5

[146] S. Mo Jones-Jang, Tara Mortensen, and Jingjing Liu. Does Media Literacy Help Identification of Fake News? Information Literacy Helps, but Other Literacies Don't. *American Behavioral Scientist*, 65(2):371–388, February 2021. ISSN 0002-7642. doi: 10.1177/0002764219869406. URL https://doi.org/10.1177/0002764219869406. 1.3.3, 2.3, 6.4.6, A.6

[147] Steffen Kallbekken, Jorge H. Garcia, and Kristine Korneliussen. Determinants of public support for transport taxes. *Transportation Research Part A: Policy and Practice*, 58: 67–78, December 2013. ISSN 09658564. doi: 10.1016/j.tra.2013.10.004. URL https://linkinghub.elsevier.com/retrieve/pii/S0965856413001845. 5.3.1, 5.5

[148] Sarawut Kankham and Jian-Ren Hou. Community Notes vs. Related Articles: Assessing Real-World Integrated Counter-Rumor Features in Response to Different Rumor Types on Social Media. *International Journal of Human–Computer Interaction*, pages 1–15, September 2024. ISSN 1044-7318. doi: 10.1080/10447318.2024.2400389. URL https://doi.org/10.1080/10447318.2024.2400389. 5.6.1, 6.4.4

[149] Joel Kaplan. More Speech and Fewer Mistakes, January 2025. URL https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/. 1.5.2, 6.2.4

[150] Hansika Kapoor, Sarah Rezaei, Swanaya Gurjar, Anirudh Tagat, Denny George, Yash Budhwar, and Arathy Puthillam. Does incentivization promote sharing "true" content online? *Harvard Kennedy School Misinformation Review*, August 2023. doi: 10.37016/mr-2020-120. URL https://misinforeview.hks.harvard.edu/article/does-incentivization-promote-sharing-true-content-online/. 7.3.2, 7.4.2

[151] Matthew Katsaros, Kathy Yang, and Lauren Fratamico. 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*, pages 477–487. Association for the Advancement of Artificial Intelligence, May 2022. doi: 10.1609/icwsm.v16i1.19308. URL https://ojs.aaai.org/index.php/ICWSM/article/view/19308. 5.1, A.1

[152] Kulvinder Kaur and Samrat Gupta. Towards dissemination, detection and combating misinformation on social media: a literature review. *Journal of Business & Industrial Marketing*, 38(8):1656–1674, January 2022. ISSN 0885-8624. doi: 10.1108/JBIM-02-2022-0066. URL https://doi.org/10.1108/JBIM-02-2022-0066. 3.1

[153] Tanveer Khan, Antonis Michalas, and Adnan Akhunzada. Fake news outbreak 2021: Can we stop the viral spread? *Journal of Network and Computer Applications*, 190:103112, September 2021. ISSN 1084-8045. doi: 10.1016/j.jnca.2021.103112. URL `https://www.sciencedirect.com/science/article/pii/S1084804521001326`. A.2

[154] Eugene Kiely and Lori Robertson. How to Spot Fake News, November 2016. URL `https://www.factcheck.org/2016/11/how-to-spot-fake-news/`. 4.3.4

[155] Jan H. Kietzmann, Kristopher Hermkens, Ian P. McCarthy, and Bruno S. Silvestre. Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons*, 54(3):241–251, May 2011. ISSN 0007-6813. doi: 10.1016/j.bushor.2011.01.005. URL `https://www.sciencedirect.com/science/article/pii/S0007681311000061`. 1.3.3, 3.4.1, 7

[156] Jisu Kim, Curtis McDonald, Paul Meosky, Matthew Katsaros, and Tom Tyler. Promoting Online Civility Through Platform Architecture. *Journal of Online Trust and Safety*, 1 (4), September 2022. ISSN 2770-3142. doi: 10.54501/jots.v1i4.54. URL `https://tsjournal.org/index.php/jots/article/view/54`. 2.3, A.1

[157] Catherine King, Christine Sowa Lepird, and Dr Kathleen M Carley. Project OMEN: Designing a Training Game to Fight Misinformation on Social Media. Technical Report CMU-ISR-21-110, Carnegie Mellon University, August 2021. URL `http://reports-archive.adm.cs.cmu.edu/anon/isr2021/abstracts/21-110.html`. 4.3, 4.3.1

[158] Catherine King, Samantha Phillips, and Kathleen Carley. Registered Report: A path forward on online misinformation mitigation based on current user behavior [Registered Report Stage 1 protocol]. *Scientific Reports*, 2024. URL `https://figshare.com/s/683b1e7c2f2bad96f604`. 4.3.3

[159] Catherine King, Samantha C. Phillips, and Kathleen M. Carley. A path forward on online misinformation mitigation based on current user behavior. *Scientific Reports*, 15(1):9475, March 2025. ISSN 2045-2322. doi: 10.1038/s41598-025-93100-7. URL `https://www.nature.com/articles/s41598-025-93100-7`. 3.1, 7.1.2, 7.4.2

[160] Jan Kirchner and Christian Reuter. 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, October 2020. ISSN 2573-0142. doi: 10.1145/3415211. URL `https://dl.acm.org/doi/10.1145/3415211`. 3.3.1, 5.1, 6.4.1, 6.4.4, 7.3.5, A.1

[161] Jason Koebler. Reddit Issuing 'Formal Legal Demands' Against Researchers Who Conducted Secret AI Experiment on Users, April 2025. URL `https://www.404media.co/reddit-issuing-formal-legal-demands-against-researchers-who-conducted-secret-ai-experiment-on-users/`. 7.3.5

[162] Marcus Kolga. Stemming the VIRUS: Understanding and responding to the threat of Russian disinformation. Technical report, Macdonald-Laurier Institute Publication, January

2019. 5.1

[163] 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`. 5

[164] Vasilis Koulolias, Gideon Mekonnen Jonathan, Miriam Fernandez, and Dimitris Sotirchos. Combating Misinformation: An ecosystem in co-creation. Technical report, OECD Publishing, January 2018. 5, 5.1, 7.4.4, 7.4.4

[165] Anastasia Kozyreva, Stephan Lewandowsky, and Ralph Hertwig. Citizens Versus the Internet: Confronting Digital Challenges With Cognitive Tools. *Psychological Science in the Public Interest*, 21(3):103–156, 2020. doi: https://doi.org/10.1177/15291006209467. URL `https://journals.sagepub.com/doi/full/10.1177/1529100620946707`. 3.3.1

[166] Anastasia Kozyreva, Stefan M. Herzog, Stephan Lewandowsky, Ralph Hertwig, Philipp Lorenz-Spreen, Mark Leiser, and Jason Reifler. Resolving content moderation dilemmas between free speech and harmful misinformation. *Proceedings of the National Academy of Sciences*, 120(7):e2210666120, February 2023. doi: 10.1073/pnas.2210666120. URL `https://www.pnas.org/doi/10.1073/pnas.2210666120`. 3.3.1, 3.3.4

[167] Anastasia Kozyreva, Philipp Lorenz-Spreen, Stefan M. Herzog, Ullrich K. H. Ecker, Stephan Lewandowsky, Ralph Hertwig, Ayesha Ali, Joe Bak-Coleman, Sarit Barzilai, Melisa Basol, Adam J. Berinsky, Cornelia Betsch, John Cook, Lisa K. Fazio, Michael Geers, Andrew M. Guess, Haifeng Huang, Horacio Larreguy, Rakoen Maertens, Folco Panizza, Gordon Pennycook, David G. Rand, Steve Rathje, Jason Reifler, Philipp Schmid, Mark D. Smith, Briony Swire-Thompson, Paula Szewach, Sander van der Linden, and Sam Wineburg. Toolbox of individual-level interventions against online misinformation. *Nature Human Behaviour*, pages 1–9, May 2024. ISSN 2397-3374. doi: 10.1038/s41562-024-01881-0. URL `https://www.nature.com/articles/s41562-024-01881-0`. 1.3.3, 1.4.1, 1.4, 1.4.1, 1.4.3, 2.1, 2.4.4, 4.1, 6.2.4, 7.1.1

[168] Michael E. Kraft and Scott R. Furlong. *Public Policy: Politics, Analysis, and Alternatives*. CQ Press, 6th edition, April 2017. ISBN 978-1-5063-5814-7. 6.1, 6.2.4

[169] John K. Kruschke and Torrin M. Liddell. The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, 25(1):178–206, February 2018. ISSN 1531-5320. doi: 10.3758/s13423-016-1221-4. URL `https://doi.org/10.3758/s13423-016-1221-4`. 3.3.2

[170] Joris Lammers, Diederik A. Stapel, and Adam D. Galinsky. Power Increases Hypocrisy: Moralizing in Reasoning, Immorality in Behavior. *Psychological Science*, 21(5):737–744, May 2010. ISSN 0956-7976. doi: 10.1177/0956797610368810. URL `https://doi.org/10.1177/0956797610368810`. 3.3.1

[171] Darren Langdridge and Trevor Butt. The fundamental attribution error: A phenomenological critique. *British Journal of Social Psychology*, 43(3):357–369, 2004. ISSN 2044-8309. doi: 10.1348/0144666042037962. URL https://onlinelibrary.wiley.com/doi/abs/10.1348/0144666042037962. 3.3.1

[172] M. Asher Lawson, Shikhar Anand, and Hemant Kakkar. Tribalism and tribulations: The social costs of not sharing fake news. *Journal of Experimental Psychology: General*, 152 (3):611–631, March 2023. ISSN 1939-2222, 0096-3445. doi: 10.1037/xge0001374. URL https://doi.apa.org/doi/10.1037/xge0001374. 3.5

[173] David Lazer and Sandra González-Bailón. Mark Zuckerberg's Immoderate Proposal, January 2025. URL https://techpolicy.press/mark-zuckerbergs-immoderate-proposal. 6.2.4

[174] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. The science of fake news. *Science*, 359(6380):1094–1096, March 2018. doi: 10.1126/science.aao2998. URL https://www.science.org/doi/full/10.1126/science.aao2998. 1, 1.2.1

[175] Michael D Lee and Eric-Jan Wagenmakers. *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press, 2013. 3.3.2, 3.3.4

[176] 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. 4.1, 4.2.1

[177] Christine Sowa Lepird. *Digital Pink Slime: Measuring, Finding, and Countering Online Threats to Local News*. PhD thesis, Carnegie Mellon University, November 2024. URL https://christine-lepird.github.io/Lepird_Thesis_Oct2024.pdf. 4.3.3, 4.3.4

[178] Jimmie Leppink, Patricia O'Sullivan, and Kal Winston. Effect size – large, medium, and small. *Perspectives on Medical Education*, 5(6):347–349, December 2016. ISSN 2212-277X. doi: 10.1007/s40037-016-0308-y. URL https://doi.org/10.1007/s40037-016-0308-y. 3.3.2

[179] Stephan Lewandowsky and John Cook. The Conspiracy Theory Handbook. Technical report, March 2020. URL http://sks.to/conspiracy. 3

[180] 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, 1.3.3, 4.2.1, A.6

[181] Stephan Lewandowsky, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest*, 13(3):106–131, December 2012. ISSN 1529-1006. doi: 10.1177/1529100612451018. 1.3.2, 1.3.2

[182] Stephan Lewandowsky, John Cook, Ullrich 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 J. Wood, and Maria S. Zaragoza. Debunking Handbook 2020, 2020. URL `https://sks.to/db2020`. 1.4.1, 3, 3.1, 4.2.2, 1, 3, A.2

[183] Michael Lipka. Americans largely foresee AI having negative effects on news, journalists, April 2025. URL `https://www.pewresearch.org/short-reads/2025/04/28/americans-largely-foresee-ai-having-negative-effects-on-news-journalists/`. 7.3.5

[184] Yi Liu, T Pinar Yildirim, and Z John Zhang. Implications of Revenue Models and Technology for Content Moderation Strategies. *SSRN Electronic Journal*, December 2021. doi: 10.2139/ssrn.3969938. URL `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3969938`. 2.5, 5.1, 7.3.3

[185] Zhuoran Lu, Patrick Li, Weilong Wang, and Ming Yin. The Effects of AI-based Credibility Indicators on the Detection and Spread of Misinformation under Social Influence. In *Proceedings of the ACM on Human-Computer Interaction*, volume 6 of *CSCW*, pages 461:1–461:27. Association for Computing Machinery, November 2022. doi: 10.1145/3555562. URL `https://dl.acm.org/doi/10.1145/3555562`. 2.3, A.2

[186] 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. 2.3, 4.2.1, 6.2.4, A.6

[187] Cameron Martel and David G. Rand. Fact-checker warning labels are effective even for those who distrust fact-checkers. *Nature Human Behaviour*, pages 1–11, September 2024. ISSN 2397-3374. doi: 10.1038/s41562-024-01973-x. URL `https://www.nature.com/articles/s41562-024-01973-x`. 7.3.4

[188] Cameron Martel, Adam J. Berinsky, David Rand, Amy Xian Zhang, and Paul Resnick. Perceived legitimacy of layperson and expert content moderators, June 2024. URL `https://osf.io/5n9tp_v1`. 6.2.4

[189] Justin D Martin and Fouad Hassan. Testing Classical Predictors of Public Willingness to Censor on the Desire to Block Fake News Online. *Convergence: The International Journal of Research into New Media Technologies*, 28(3):867–887, June 2022. ISSN 1354-8565, 1748-7382. doi: 10.1177/13548565211012552. URL `https://journals.sagepub.com/doi/10.1177/13548565211012552`. 5.2

[190] Francisco J. Martínez-López, Yangchun Li, and Susan M. Young. Social Media Monetization and Demonetization: Risks, Challenges, and Potential Solutions. In Francisco J. Martínez-López, Yangchun Li, and Susan M. Young, editors, *Social Media Monetization: Platforms, Strategic Models and Critical Success Factors*, pages 185–214. Springer International Publishing, Cham, 2022. ISBN 978-3-031-14575-9. doi: 10.1007/978-3-031-14575-9_13. URL `https://doi.org/10.1007/978-3-031-14575-9_13`. 2.5,

6.4.3, A.3

[191] Stefan D. McCabe, Diogo Ferrari, Jon Green, David M. J. Lazer, and Kevin M. Esterling. Post-January 6th deplatforming reduced the reach of misinformation on Twitter. *Nature*, 630(8015):132–140, June 2024. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-024-07524-8. URL https://www.nature.com/articles/s41586-024-07524-8. 6.4.3

[192] Colleen McClain, Brian Kennedy, Jeffrey Gottfried, Monica Anderson, and Giancarlo Pasquini. How the U.S. Public and AI Experts View Artificial Intelligence. Technical report, Pew Research Center, April 2025. URL https://www.pewresearch.org/internet/2025/04/03/how-the-us-public-and-ai-experts-view-artificial-intelligence/. 7.3.5, 7.4.3

[193] Sarah McGrew. Learning to evaluate: An intervention in civic online reasoning. *Computers & Education*, 145:103711, February 2020. ISSN 0360-1315. doi: 10.1016/j.compedu.2019.103711. 4.1

[194] Colten Meisner. The weaponization of platform governance: Mass reporting and algorithmic punishments in the creator economy. *Policy & Internet*, 15(4):466–477, 2023. ISSN 1944-2866. doi: 10.1002/poi3.359. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/poi3.359. 7.3.4

[195] Gianna Melillo. Nearly half of Americans say they see misinformation on a daily basis: survey. *The Hill*, August 2022. URL https://thehill.com/changing-america/enrichment/education/3597328-nearly-half-of-americans-say-they-see-misinformation-on-a-daily-basis-survey/. 1

[196] Paul Mena. Cleaning Up Social Media: The Effect of Warning Labels on Likelihood of Sharing False News on Facebook. *Policy & Internet*, 12(2):165–183, 2020. ISSN 1944-2866. doi: 10.1002/poi3.214. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/poi3.214. 5.1, 6.4.4, 7.3.5

[197] Trevor van Mierlo. The 1% Rule in Four Digital Health Social Networks: An Observational Study. *Journal of Medical Internet Research*, 16(2):e2966, February 2014. doi: 10.2196/jmir.2966. URL https://www.jmir.org/2014/2/e33. 7.4.1

[198] Susan Minichiello. California now has a law to bolster media literacy in schools, September 2018. URL https://www.pressdemocrat.com/article/news/california-now-has-a-law-to-bolster-media-literacy-in-schools/. 7.4.3

[199] Amy Mitchell and Mason Walker. More Americans now say government should take steps to restrict false information online than in 2018, 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/. 3.3.1, 3.3.4, 5.2

[200] Ariana Modirrousta-Galian and Philip A. Higham. 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, 2023. ISSN 1939-2222. doi: 10.1037/xge0001395. 4.2.1, A.6

[201] Richard D. Morey, Jeffrey N. Rouder, Tahira Jamil, Urbanek, Karl Forner, and Alexander Ly. Package ‘BayesFactor’, 2022. URL `https://richarddmorey.github.io/BayesFactor/`. 3.3.2, 3.3.3

[202] 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. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.24637`. A.4

[203] Mohsen Mosleh, Jennifer Nancy Lee Allen, and David Rand. Divergent patterns of engagement with partisan and low-quality news across seven social media platforms, December 2024. URL `https://osf.io/9csy3`. 3.4.1, 3.4.1

[204] Mohsen Mosleh, Qi Yang, Tauhid Zaman, Gordon Pennycook, and David G. Rand. Differences in misinformation sharing can lead to politically asymmetric sanctions. *Nature*, pages 1–8, October 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07942-8. URL `https://www.nature.com/articles/s41586-024-07942-8`. 7.3.3

[205] Joris Mulder, Donald R. Williams, Xin Gu, Andrew Tomarken, Florian Böing-Messing, Anton Olsson-Collentine, Marlyne Meijerink, Janosch Menke, Robbie Van Aert, Jean-Paul Fox, Herbert Hoijtink, Yves Rosseel, Eric-Jan Wagenmakers, and Caspar Van Lissa. **BFpack** : Flexible Bayes Factor Testing of Scientific Theories in *R*. *Journal of Statistical Software*, 100(18), 2021. ISSN 1548-7660. doi: 10.18637/jss.v100.i18. URL `https://www.jstatsoft.org/v100/i18/`. 5.3.3

[206] Conor Murray. Avril Lavigne Clone Conspiracy Explained: Singer Laughs Off False Rumor—Here’s How It All Began, May 2024. URL `https://www.forbes.com/sites/conormurray/2024/05/16/avril-lavigne-clone-conspiracy-explained-singer-laughs-off-false-rumor-heres-how-it-all-began/`. 1.3.2

[207] Xiaoli Nan, Yuan Wang, and Kathryn Thier. Why do people believe health misinformation and who is at risk? A systematic review of individual differences in susceptibility to health misinformation. *Social Science & Medicine*, 314:115398, December 2022. ISSN 0277-9536. doi: 10.1016/j.socscimed.2022.115398. URL `https://www.sciencedirect.com/science/article/pii/S0277953622007043`. 3.3.1, 3.5

[208] Jack Nassetta and Kimberly Gross. State media warning labels can counteract the effects of foreign disinformation. *Harvard Kennedy School Misinformation Review*, 1(Special Issue on US Elections and Disinformation), October 2020. doi: 10.37016/mr-2020-45. URL `https://misinforeview.hks.harvard.edu/article/state-media-warning-labels-can-counteract-the-effects-of-foreign-misinformation/`. 2.3, 6.4.4, A.4

[209] 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://doi.org/10.17226/25335. 1.2.1, 7

[210] Andrew A. Neath, Javier E. Flores, and Joseph E. Cavanaugh. Bayesian multiple comparisons and model selection. *WIREs Computational Statistics*, 10(2):e1420, 2018. ISSN 1939-0068. doi: 10.1002/wics.1420. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.1420. 3.3.2

[211] Casey Newton. Google gives up on data voids, February 2025. URL https://www.platformer.news/google-data-voids-warning-banners-2024-election/. 7.3.3

[212] Lynnette H. X. Ng and Araz Taeihagh. How does fake news spread? Understanding pathways of disinformation spread through APIs. *Policy & Internet*, 13 (4):560–585, 2021. ISSN 1944-2866. doi: 10.1002/poi3.268. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/poi3.268. 1.3.2, 1.3.3, 1.3.3, 1.3.3, 7.1.1

[213] Sheryl Wei Ting Ng. Self- and Social Corrections on Instant Messaging Platforms. *International Journal Of Communication*, 17:426–446, 2023. URL https://ijoc.org/index.php/ijoc/article/view/19142/4005. 3.3.1, 3.4.1, 1

[214] 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. 4.2.2, 7.4.3, A.5, A.7

[215] Brendan Nyhan and Jason Reifler. When Corrections Fail: The Persistence of Political Misperceptions. *Political Behavior*, 32(2):303–330, 2010. ISSN 0190-9320. URL https://www.jstor.org/stable/40587320. 7.3.4

[216] Katherine Ognyanova, David Lazer, Ronald E. Robertson, and Christo Wilson. Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. *Harvard Kennedy School Misinformation Review*, June 2020. doi: 10.37016/mr-2020-024. URL https://misinforeview.hks.harvard.edu/article/misinformation-in-action-fake-news-exposure-is-linked-to-lower-trust-in-media-higher-trust-in-government-when-your-side-is-in-power/. 6.4.7

[217] Chitu Okoli. A Guide to Conducting a Standalone Systematic Literature Review. *Communications of the Association for Information Systems*, 37, 2015. ISSN 15293181. doi: 10.17705/1CAIS.03743. URL https://aisel.aisnet.org/cais/vol37/iss1/43/. 2.2, 2.2.4, 7

[218] 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

[219] Gábor Orosz, Péter Krekó, Benedek Paskuj, István Tóth-Király, Beáta Bőthe, and Christine Roland-Lévy. Changing Conspiracy Beliefs through Rationality and Ridiculing. *Frontiers in Psychology*, 7, October 2016. ISSN 1664-1078. doi: 10.3389/fpsyg.2016.01525. URL https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2016.01525/full. 3

[220] Wei Pan. Akaike's Information Criterion in Generalized Estimating Equations. *Biometrics*, 57(1):120–125, March 2001. ISSN 0006-341X. doi: 10.1111/j.0006-341X.2001.00120.x. URL https://doi.org/10.1111/j.0006-341X.2001.00120.x. D

[221] Emmanouil Papadogiannakis, Panagiotis Papadopoulos, Evangelos P. Markatos, and Nicolas Kourtellis. Who Funds Misinformation? A Systematic Analysis of the Ad-related Profit Routines of Fake News Sites. In *Proceedings of the ACM Web Conference 2023*, pages 2765–2776, Austin TX USA, April 2023. ACM. ISBN 978-1-4503-9416-1. doi: 10.1145/3543507.3583443. URL https://dl.acm.org/doi/10.1145/3543507.3583443. 2.5

[222] Orestis Papakyriakopoulos and Ellen Goodman. 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, pages 2541–2551, New York, NY, USA, April 2022. Association for Computing Machinery. ISBN 978-1-4503-9096-5. doi: 10.1145/3485447.3512126. URL https://doi.org/10.1145/3485447.3512126. 5.1, A.4

[223] Christopher V. Pece and Gary W. Anderson. Analysis of Federal Funding for Research and Development in 2022: Basic Research. Technical Report NSF 24-332, National Center for Science and Engineering, August 2024. URL https://ncses.nsf.gov/pubs/nsf24332. 7.4.4

[224] Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*, 46(4): 1023–1031, December 2014. ISSN 1554-3528. doi: 10.3758/s13428-013-0434-y. URL https://doi.org/10.3758/s13428-013-0434-y. 3.2.3

[225] Mark J. Pelletier, Alexandra Krallman, Frank G. Adams, and Tyler Hancock. One size doesn't fit all: a uses and gratifications analysis of social media platforms. *Journal of Research in Interactive Marketing*, 14(2):269–284, June 2020. ISSN 2040-7122. doi: 10.1108/JRIM-10-2019-0159. URL https://www.emerald.com/insight/content/doi/10.1108/jrim-10-2019-0159/full/html. 3.4.1

[226] 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. 3.5, 7.5

[227] Gordon Pennycook and David G. Rand. Accuracy prompts are a replicable and gener-

alizable approach for reducing the spread of misinformation. *Nature Communications*, 13(1):2333, April 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-30073-5. URL `https://www.nature.com/articles/s41467-022-30073-5`. 6.4.1

[228] Gordon Pennycook and David G. Rand. Nudging Social Media toward Accuracy. *The ANNALS of the American Academy of Political and Social Science*, 700(1):152–164, March 2022. ISSN 0002-7162, 1552-3349. doi: 10.1177/00027162221092342. URL `https://journals.sagepub.com/doi/10.1177/00027162221092342`. 5.1, 5.5

[229] 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. URL `https://pubsonline.informs.org/doi/abs/10.1287/mnsc.2019.3478`. 7.3.5

[230] Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G. Lu, and David G. Rand. Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention. *Psychological Science*, 31(7):770–780, July 2020. ISSN 0956-7976. doi: 10.1177/0956797620939054. URL `https://doi.org/10.1177/0956797620939054`. 3.3.2

[231] 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. URL `https://www.nature.com/articles/s41586-021-03344-2`. 3.1, 3.5, 5.1, A.1

[232] Paul K. Piff, Michael W. Kraus, Stéphane Côté, Bonnie Hayden Cheng, and Dacher Keltner. Having less, giving more: The influence of social class on prosocial behavior. *Journal of Personality and Social Psychology*, 99(5):771–784, 2010. ISSN 1939-1315, 0022-3514. doi: 10.1037/a0020092. URL `https://doi.apa.org/doi/10.1037/a0020092`. 3.3.5

[233] Mark Aaron Polger. Misinformation and Disinformation: Thinking Critically about Information Sources, October 2024. URL `https://library.csi.cuny.edu/misinformation/spotfakenews`. 4.3.4

[234] Ethan Porter and Thomas J. Wood. The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom. *Proceedings of the National Academy of Sciences*, 118(37):e2104235118, September 2021. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2104235118. URL `https://pnas.org/doi/full/10.1073/pnas.2104235118`. 1, 6.2.4, 7.3.4

[235] Jon Porter. WhatsApp says its forwarding limits have cut the spread of viral messages by 70 percent, April 2020. URL `https://www.theverge.com/2020/4/27/21238082/whatsapp-forward-message-limits-viral-misinformation-decline`. 6.4.1, A.1

[236] Carlie Porterfield. Twitter Begins Asking Users To Actually Read Articles Before Sharing Them, June 2020. URL `https://www.forbes.com/sites/`

carlieporterfield/2020/06/10/twitter-begins-asking-users-to-actually-read-articles-before-sharing-them/. 5.1, 5.5, A.1

[237] Elaine S. Povich. Internet Ads Are a Popular Tax Target for Both Parties, June 2021. URL https://pew.org/3psT1Rq. A.7

[238] Roxana Radu. Fighting the 'Infodemic': Legal Responses to COVID-19 Disinformation. *Social Media + Society*, 6(3):2056305120948190, July 2020. ISSN 2056-3051. doi: 10.1177/2056305120948190. URL https://doi.org/10.1177/2056305120948190. 5.1, 5.2

[239] David Rand and Cameron Martel. Americans actually do want expert content moderation, January 2025. URL https://thehill.com/opinion/technology/5109667-sorry-zuckerberg-americans-actually-do-want-expert-content-moderation/. 6.2.4

[240] 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. *Perspectives on Psychological Science*, 19(5):781–795, September 2024. ISSN 1745-6916. doi: 10.1177/17456916231190392. URL https://doi.org/10.1177/17456916231190392. 7.4.2

[241] Adrian Rauchfleisch and Jonas Kaiser. The impact of deplatforming the far right: an analysis of YouTube and BitChute. *Information, Communication & Society*, 27(7):1478–1496, May 2024. ISSN 1369-118X, 1468-4462. doi: 10.1080/1369118X.2024.2346524. URL https://www.tandfonline.com/doi/full/10.1080/1369118X.2024.2346524. 5.1, 6.4.3, 7.3.4, A.3

[242] Emma Remy. How public and private Twitter users in the U.S. compare — and why it might matter for your research, July 2019. URL https://www.pewresearch.org/decoded/2019/07/how-public-and-private-twitter-users-in-the-u-s-compare-and-why-it-might-matter-for-your-research/. 3.4.1

[243] J. P. Reynolds, K. Stautz, M. Pilling, S. van der Linden, and T. M. Marteau. Communicating the effectiveness and ineffectiveness of government policies and their impact on public support: a systematic review with meta-analysis. *Royal Society Open Science*, 7(1):190522, January 2020. doi: 10.1098/rsos.190522. URL https://doi.org/10.1098/rsos.190522. 5.5, 6.2.3

[244] 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. 5.2

[245] Diana Ridley. *The Literature Review: A Step-by-Step Guide for Students*. SAGE Publications Ltd, Los Angeles London New Delhi, second edition edition, July 2012. ISBN 978-1-4462-0142-8. 2.2.4

[246] Ronald E. Robertson, Evan M. Williams, Kathleen M. Carley, and David Thiel. Data Voids and Warning Banners on Google Search, February 2025. URL http://arxiv.org/

abs/2502.17542. 7.3.3

[247] Alexander Robitzsch. Why Ordinal Variables Can (Almost) Always Be Treated as Continuous Variables: Clarifying Assumptions of Robust Continuous and Ordinal Factor Analysis Estimation Methods. *Frontiers in Education*, 5, 2020. ISSN 2504-284X. URL https://www.frontiersin.org/articles/10.3389/feduc.2020.589965. 3.3.3

[248] 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. URL https://www.tandfonline.com/doi/full/10.1080/10811680.2020.1735194. 5, 6.1, 6.2.1, 6.2.4, 7, 7.1.1, 7.4.3, 7.4.3, A.7

[249] 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. 4.2.1

[250] 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, 2.5, A.6

[251] Jon Roozenbeek, Sander van der Linden, Beth Goldberg, Steve Rathje, and Stephan Lewandowsky. Psychological inoculation improves resilience against misinformation on social media. *Science Advances*, 8(34):eabo6254, August 2022. doi: 10.1126/sciadv.abo6254. URL https://www.science.org/doi/10.1126/sciadv.abo6254. 5.1

[252] Bettina Rottweiler and Paul Gill. Conspiracy Beliefs and Violent Extremist Intentions: The Contingent Effects of Self-efficacy, Self-control and Law-related Morality. *Terrorism and Political Violence*, 34(7):1485–1504, October 2020. ISSN 0954-6553. doi: 10.1080/09546553.2020.1803288. URL https://doi.org/10.1080/09546553.2020.1803288. 5.1

[253] Jeffrey N. Rouder, Paul L. Speckman, Dongchu Sun, Richard D. Morey, and Geoffrey Iverson. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16(2):225–237, April 2009. ISSN 1069-9384, 1531-5320. doi: 10.3758/PBR.16.2.225. URL http://link.springer.com/10.3758/PBR.16.2.225. 3.3.2

[254] Johnny Saldaña. *The Coding Manual for Qualitative Researchers*. Sage, 2nd edition, 2013. ISBN 978-1-4462-4737-2. 4.3.5

[255] Emily Saltz, Soubhik Barari, Claire 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. URL https://misinforeview.hks.harvard.edu/article/misinformation-interventions-are-common-divisive-and-poorly-understood/. 3.1, 3.3.1, 3.3.4, 3.3.5, 5.2, 5.5, 6.2.3, 6.2.3, 7.4.2

[256] Sydney Schaedel. Did the Pope Endorse Trump?, October 2016. URL https://www.factcheck.org/2016/10/did-the-pope-endorse-trump/. 1.3.2

[257] Brennan Schaffner, Arjun Nitin Bhagoji, Siyuan Cheng, Jacqueline Mei, Jay L. Shen, Grace Wang, Marshini Chetty, Nick Feamster, Genevieve Lakier, and Chenhao Tan. "Community Guidelines Make this the Best Party on the Internet": An In-Depth Study of Online Platforms' Content Moderation Policies. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–16, May 2024. doi: 10.1145/3613904.3642333. URL http://arxiv.org/abs/2405.05225. 3.4.1, A.2

[258] Mark Schlesinger and Caroline Heldman. Gender Gap or Gender Gaps? New Perspectives on Support for Government Action and Policies. *Journal of Politics*, 63 (1):59–92, 2001. ISSN 1468-2508. doi: 10.1111/0022-3816.00059. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/0022-3816.00059. 5.5

[259] F. D. Schönbrodt and A. M. Stefan. BFDA: An R package for Bayes factor design analysis (version 0.5.0), 2019. URL https://github.com/nicebread/BFDA. 3.3.2

[260] Felix D. Schönbrodt and Eric-Jan Wagenmakers. Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, 25(1):128–142, February 2018. ISSN 1531-5320. doi: 10.3758/s13423-017-1230-y. URL https://doi.org/10.3758/s13423-017-1230-y. 3.3.2

[261] Amanda Seitz and Hannah Fingerhut. Americans agree misinformation is a problem, poll shows, October 2021. URL https://apnews.com/article/coronavirus-pandemic-technology-business-health-misinformation-fbe9d09024d7b92e1600e411d5f931dd. 3.3.2

[262] Filipo Sharevski, Raniem Alsaadi, Peter Jachim, and Emma Pieroni. Misinformation warnings: Twitter's soft moderation effects on COVID-19 vaccine belief echoes. *Computers & Security*, 114:102577, March 2022. ISSN 0167-4048. doi: 10.1016/j.cose.2021.102577. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8675217/. 2.3, 6.4.1, 6.4.4, A.1, A.4

[263] Olivia Sidoti and Wyatt Dawson. Social Media Fact Sheet, November 2024. URL https://www.pewresearch.org/internet/fact-sheet/social-media/. 1.5.1

[264] Sean Simpson. Fake News: A Global Epidemic Vast Majority (86%) of Online Global Citizens Have Been Exposed to it. Technical report, Ipsos, Toronto, Canada, June 2019. URL https://www.ipsos.com/en-us/news-polls/cigi-fake-news-global-epidemic. 3.3.2

[265] Ciarra N. Smith and Holli H. Seitz. Correcting Misinformation About Neuroscience via Social Media. *Science Communication*, 41(6):790–819, December 2019. ISSN 1075-5470. doi: 10.1177/1075547019890073. URL https://doi.org/10.1177/1075547019890073. A.1

[266] Leanne S. Son Hing, Winnie Li, and Mark P. Zanna. Inducing Hypocrisy to Reduce Prejudicial Responses among Aversive Racists. *Journal of Experimental Social Psychology*, 38(1):71–78, January 2002. ISSN 0022-1031. doi:

10.1006/jesp.2001.1484. URL https://www.sciencedirect.com/science/article/pii/S0022103101914842. 7.3.1

[267] Brian G. Southwell, Jeff Niederdeppe, Joseph N. Cappella, Anna Gaysynsky, Dannielle E. Kelley, April Oh, Emily B. Peterson, and Wen-Ying Sylvia Chou. Misinformation as a Misunderstood Challenge to Public Health. *American Journal of Preventive Medicine*, 57(2):282–285, August 2019. ISSN 07493797. doi: 10.1016/j.amepre.2019.03.009. URL https://linkinghub.elsevier.com/retrieve/pii/S074937971930159X. 3.3.1, 1, 7.3.1

[268] Saranac Hale Spencer. Fake Coronavirus Cures, Part 2: Garlic Isn't a 'Cure', February 2020. URL https://www.factcheck.org/2020/02/fake-coronavirus-cures-part-2-garlic-isnt-a-cure/. 1.3.2

[269] Christopher St. Aubin and Jacob Liedke. Social Media and News Fact Sheet. Technical report, Pew Research Center, September 2024. URL https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/. 1

[270] 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. 3.1, 3.5, 7.3.1

[271] Jeff Stone and Nicholas C. Fernandez. When thinking about less failure causes more dissonance: The effect of elaboration and recall on behavior change following hypocrisy. *Social Influence*, 6(4):199–211, October 2011. ISSN 1553-4510. doi: 10.1080/15534510.2011.618368. URL https://doi.org/10.1080/15534510.2011.618368. 7.3.1

[272] Victor Suarez-Lledo and Javier Alvarez-Galvez. Prevalence of Health Misinformation on Social Media: Systematic Review. *Journal of Medical Internet Research*, 23(1): e17187, January 2021. doi: 10.2196/17187. URL https://www.jmir.org/2021/1/e17187. 3.4.1

[273] Edson C. Tandoc, Zheng Lim, and Rich Ling. Defining "Fake News": A typology of scholarly definitions. *Digital Journalism*, 6:1–17, August 2017. doi: 10.1080/21670811.2017.1360143. 1.2.2, 1.1, 1.3.1

[274] 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. 3.3.1, 3.4.1, 1, 7.3.1, A.5

[275] 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. 1.4.1, 4.2.2

[276] Josh Taylor. Threads: how do I sign up and is it any different to Twitter? *The Guardian*, July 2023. ISSN 0261-3077. URL https://www.theguardian.com/

technology/2023/jul/06/threads-how-do-i-sign-up-and-is-it-any-different-to-twitter. 3.4.1

[277] Michael Henry Tessler, Michiel A. Bakker, Daniel Jarrett, Hannah Sheahan, Martin J. Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C. Parkes, Matthew Botvinick, and Christopher Summerfield. AI can help humans find common ground in democratic deliberation. *Science*, 386(6719):eadq2852, October 2024. doi: 10.1126/science.adq2852. URL https://www.science.org/doi/10.1126/science.adq2852. 7.3.5

[278] Philip E. Tetlock. Accountability: A Social Check on the Fundamental Attribution Error. *Social Psychology Quarterly*, 48(3):227–236, 1985. ISSN 0190-2725. doi: 10.2307/3033683. URL https://www.jstor.org/stable/3033683. 3.3.1

[279] David Thiel, Renée DiResta, and Alex Stamos. Cross-Platform Dynamics of Self-Generated CSAM. Technical report, Stanford Internet Observatory, June 2023. 7.3.3, 7.4.2

[280] Daniel Robert Thomas and Laila A. Wahedi. Disrupting hate: The effect of deplatforming hate organizations on their online audience. *Proceedings of the National Academy of Sciences*, 120(24):e2214080120, June 2023. doi: 10.1073/pnas.2214080120. URL https://www.pnas.org/doi/10.1073/pnas.2214080120. 2.3, 6.4.3, A.3

[281] 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. URL https://doi.org/10.1080/21670811.2021.1977668. A.7

[282] Samuel P. Trethewey. Strategies to combat medical misinformation on social media. *Postgraduate Medical Journal*, 96(1131):4–6, January 2020. ISSN 0032-5473, 1469-0756. doi: 10.1136/postgradmedj-2019-137201. URL https://pmj.bmj.com/content/96/1131/4. 3

[283] Andrea C. Tricco, Erin Lillie, Wasifa Zarin, Kelly K. O'Brien, Heather Colquhoun, Danielle Levac, David Moher, Micah D. J. Peters, Tanya Horsley, Laura Weeks, Susanne Hempel, Elie A. Akl, Christine Chang, Jessie McGowan, Lesley Stewart, Lisa Hartling, Adrian Aldcroft, Michael G. Wilson, Chantelle Garritty, Simon Lewin, Christina M. Godfrey, Marilyn T. Macdonald, Etienne V. Langlois, Karla Soares-Weiser, Jo Moriarty, Tammy Clifford, Özge Tunçalp, and Sharon E. Straus. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Annals of Internal Medicine*, 169 (7):467–473, October 2018. ISSN 1539-3704. doi: 10.7326/M18-0850. 2.2

[284] Gleb Tsipursky, Fabio Votta, and Kathryn M. Roose. Fighting Fake News and Post-Truth Politics with Behavioral Science: The Pro-Truth Pledge. *Behavior and Social Issues*, 27 (1):47–70, May 2018. ISSN 2376-6786. doi: 10.5210/bsi.v27i0.9127. URL https://doi.org/10.5210/bsi.v27i0.9127. 7.3.1, 7.4.1, 7.5

[285] Fangjing Tu. Empowering social media users: nudge toward self-engaged verification for improved truth and sharing discernment. *Journal of Communication*, 74(3):225–236,

June 2024. ISSN 0021-9916. doi: 10.1093/joc/jqae007. URL https://doi.org/10.1093/joc/jqae007. 7.3.1, 7.5

[286] Joshua Tucker, Andrew Guess, Pablo Barbera, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature. *SSRN Electronic Journal*, 2018. ISSN 1556-5068. doi: 10.2139/ssrn.3144139. URL https://www.ssrn.com/abstract=3144139. 5.1

[287] 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

[288] Melissa Tully, Emily K. Vraga, and Leticia Bode. Designing and Testing News Literacy Messages for Social Media. *Mass Communication and Society*, 23(1):22–46, January 2020. ISSN 1520-5436. doi: 10.1080/15205436.2019.1604970. URL https://doi.org/10.1080/15205436.2019.1604970. 2.3, A.6

[289] Tamara Uhaze. Research Subject Guides: Fake News/Misinformation/Disinformation: What is Fake News? URL https://subjectguides.lib.neu.edu/fakenews/. 1.3.1

[290] Jay J. Van Bavel, Steve Rathje, Madalina Vlasceanu, and Clara Pretus. Updating the identity-based model of belief: From false belief to the spread of misinformation. *Current Opinion in Psychology*, 56:101787, April 2024. ISSN 2352-250X. doi: 10.1016/j.copsyc.2023.101787. URL https://www.sciencedirect.com/science/article/pii/S2352250X23002324. 7.3.2, 7.4.2, 7.5

[291] 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. 2.6.2, 7

[292] Tiago Ventura, Rajeshwari Majumdar, Jonathan Nagler, and Joshua A. Tucker. Misinformation Exposure Beyond Traditional Feeds: Evidence from a WhatsApp Deactivation Experiment in Brazil, May 2023. URL https://papers.ssrn.com/abstract=4457400. A.5

[293] Santosh Vijaykumar, Yan Jin, Daniel Rogerson, Xuerong Lu, Swati Sharma, Anna Maughan, Bianca Fadel, Mariella Silva de Oliveira Costa, Claudia Pagliari, and Daniel Morris. How shades of truth and age affect responses to COVID-19 (Mis)information: randomized survey experiment among WhatsApp users in UK and Brazil. *Humanities and Social Sciences Communications*, 8(1):1–12, March 2021. ISSN 2662-9992. doi: 10.1057/s41599-021-00752-7. URL https://www.nature.com/articles/s41599-021-00752-7. 3.3.1

[294] Kristina De Voe. Research Guides: ”Fake News,” Misinformation & Disinformation: What is fake news? URL https://guides.temple.edu/fakenews. 1.3.1

[295] Emily A. Vogels. Support for more regulation of tech companies has declined in U.S., especially among Republicans, May 2022. URL https://www.pewresearch.org/short-reads/2022/05/13/support-for-more-regulation-of-tech-companies-has-declined-in-u-s-especially-among-republicans/. 5, 5.5, 6.2.3, 7.3.3, 7.4.3

[296] Hilde A. M. Voorveld, Guda van Noort, Daniël G. Muntinga, and Fred Bronner. Engagement with Social Media and Social Media Advertising: The Differentiating Role of Platform Type. *Journal of Advertising*, 47(1):38–54, January 2018. ISSN 0091-3367. doi: 10.1080/00913367.2017.1405754. URL https://doi.org/10.1080/00913367.2017.1405754. (document), 3.4.1, 3.4.1, 3.4.1, 3.9, 3.4.1

[297] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, March 2018. doi: 10.1126/science.aap9559. URL https://www.science.org/doi/10.1126/science.aap9559. 1, 1.2.1

[298] Emily K. Vraga, Melissa Tully, and Leticia Bode. Assessing the relative merits of news literacy and corrections in responding to misinformation on Twitter. *New Media & Society*, 24(10):2354–2371, October 2022. ISSN 1461-4448. doi: 10.1177/1461444821998691. URL https://doi.org/10.1177/1461444821998691. 7.3.4

[299] Kurt Wagner. Inside Twitter's Plan to Fact-Check Tweets. *Bloomberg.com*, March 2021. URL https://www.bloomberg.com/news/newsletters/2021-03-04/birdwatch-inside-twitter-s-plan-to-fact-check-tweets. 3.1

[300] 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. 1, 1.3.2, 1.3.3, A.2

[301] Nathan Walter and Nikita A. Salovich. Unchecked vs. Uncheckable: How Opinion-Based Claims Can Impede Corrections of Misinformation. *Mass Communication and Society*, 24 (4):500–526, July 2021. ISSN 1520-5436. doi: 10.1080/15205436.2020.1864406. URL https://doi.org/10.1080/15205436.2020.1864406. 2.3

[302] Nathan Walter, Jonathan Cohen, R. Lance Holbert, and Yasmin Morag. Fact-Checking: A Meta-Analysis of What Works and for Whom. *Political Communication*, 37(3):350–375, May 2020. ISSN 1058-4609. doi: 10.1080/10584609.2019.1668894. URL https://doi.org/10.1080/10584609.2019.1668894. 6.4.1

[303] 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. 1.4.1, 3, 4.2.2

[304] Min Wang, Mingke Rao, and Zhipeng Sun. Typology, Etiology, and Fact-Checking: A Pathological Study of Top Fake News in China. *Journalism Practice*, 16:1–19, August 2020. doi: 10.1080/17512786.2020.1806723. 1.1, 1.3.1

[305] Claire Wardle and Hossein Derakhshan. Information Disorder: Toward an interdisciplinary framework for research and policy making. Technical report, Council of Europe, September 2017. URL https://rm.coe.int/information-

disorder-toward-an-interdisciplinary-framework-for-researc/168076277c. 1.2.2, 1.3.1, 1.1, 1.3.1, 1.3.2, 1.3.2, 1.3.2, 1.3.2, 1.3.3, 1.3.3, 1.3.3, 7.1.1, 7.4.3, 7.4.4, 7.4.4, A.7

[306] 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. 1

[307] Amy Watson. Sharing of made-up news on social networks in the U.S. 2020, June 2022. URL https://www.statista.com/statistics/657111/fake-news-sharing-online/. 3.1, 3.3.2

[308] Sarah Myers West. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, 20(11):4366–4383, November 2018. ISSN 1461-4448. doi: 10.1177/1461444818773059. URL https://doi.org/10.1177/1461444818773059. 6.4.3, A.3

[309] Chloe Wittenberg, Ziv Epstein, Adam J. Berinsky, and David G. Rand. Labeling AI-Generated Content: Promises, Perils, and Future Directions. *An MIT Exploration of Generative AI*, March 2024. ISSN ,. doi: 10.21428/e4baedd9.0319e3a6. URL https://mit-genai.pubpub.org/pub/hu71se89/release/1. 7.3.5, 7.5

[310] Randy Yee Man Wong, Christy M. K. Cheung, Bo Xiao, and Jason Bennett Thatcher. Standing Up or Standing By: Understanding Bystanders' Proactive Reporting Responses to Social Media Harassment. *Information Systems Research*, 32(2):561–581, June 2021. ISSN 1047-7047. doi: 10.1287/isre.2020.0983. URL https://pubsonline.informs.org/doi/abs/10.1287/isre.2020.0983. 3.5

[311] 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. 7.3.4

[312] Huiping Wu and Shing-On Leung. Can Likert Scales be Treated as Interval Scales?—A Simulation Study. *Journal of Social Service Research*, 43(4):527–532, August 2017. ISSN 0148-8376. doi: 10.1080/01488376.2017.1329775. URL https://doi.org/10.1080/01488376.2017.1329775. 3.3.3

[313] Liang Wu, Fred Morstatter, Kathleen M. Carley, and Huan Liu. Misinformation in Social Media: Definition, Manipulation, and Detection. *ACM SIGKDD Explorations Newsletter*, 21(2):80–90, November 2019. ISSN 1931-0145, 1931-0153. doi: 10.1145/3373464.3373475. URL https://dl.acm.org/doi/10.1145/3373464.3373475. 1.2.2

[314] 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. 5, 5.1, 5.2.1, A.7

[315] 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`. 2.5, 5, 5.1, A.1, A.4

[316] Juanita Zainudin, Nazlena Mohamad Ali, Alan F. Smeaton, and Mohamad Taha Ijab. Intervention Strategies for Misinformation Sharing on Social Media: A Bibliometric Analysis. *IEEE Access*, 12:140359–140379, 2024. ISSN 2169-3536. doi: 10.1109/ACCESS.2024.3469248. URL `https://ieeexplore.ieee.org/document/10697164`. 2.2

[317] Savvas Zannettou, Michael Sirivianos, Jeremy Blackburn, and Nicolas Kourtellis. The Web of False Information: Rumors, Fake News, Hoaxes, Clickbait, and Various Other Shenanigans. *Journal of Data and Information Quality*, 11(3):1–37, September 2019. ISSN 1936-1955, 1936-1963. doi: 10.1145/3309699. URL `https://dl.acm.org/doi/10.1145/3309699`. 1.1, 1.3.1, 1.3.2, 7.1.1

[318] Maxwell Zeff. OpenAI used this subreddit to test AI persuasion, January 2025. URL `https://techcrunch.com/2025/01/31/openai-used-this-subreddit-to-test-ai-persuasion/`. 7.3.5

[319] Max Zhan. Here's why Meta ended fact-checking, according to experts, January 2025. URL `https://abcnews.go.com/US/why-did-meta-remove-fact-checkers-experts-explain/story?id=117417445`. 6.2.4

[320] Xichen Zhang and Ali A. Ghorbani. An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, 57(2):102025, March 2020. ISSN 03064573. doi: 10.1016/j.ipm.2019.03.004. URL `https://linkinghub.elsevier.com/retrieve/pii/S0306457318306794`. 1, 1.2.1, 1.3.2, 1.3.2, 1.3.2

[321] Hong Zhou, Yaobin Lu, Ling Zhao, Bin Wang, and Ting Li. 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, October 2024. ISSN 0144-929X. doi: 10.1080/0144929X.2023.2281491. URL `https://doi.org/10.1080/0144929X.2023.2281491`. 3.5, 6.2.3, 6.4.5, A.5

[322] Yu-Qian Zhu and Houn-Gee Chen. Social media and human need satisfaction: Implications for social media marketing. *Business Horizons*, 58(3):335–345, May 2015. ISSN 0007-6813. doi: 10.1016/j.bushor.2015.01.006. URL `https://www.sciencedirect.com/science/article/pii/S0007681315000075`. (document), 1.3.3, 3.4.1, 3.4.1, 3.7