



PhD-FSTM-2026-038
The Faculty of Science, Technology and Medicine

DISSERTATION

Defence held on 20 April 2026 in Luxembourg

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG EN INFORMATIQUE

by

Yuwei CHUAI

COMPUTATIONAL ANALYSIS OF MISINFORMATION ENGAGEMENT AND INTERVENTIONS ON SOCIAL MEDIA

Dissertation defence committee

Dr Gabriele LENZINI, dissertation supervisor
Associate Professor, Université du Luxembourg

Dr Jun PANG, chairman
Assistant Professor, Université du Luxembourg

Dr Martin THEOBALD, member
Professor, Université du Luxembourg

Dr Mohsen MOSLEH, member
Associate Professor, University of Oxford

Dr Xin YI, member
Associate Professor, Tsinghua University

Affidavit / Statement of originality

I declare that this thesis:

- is the result of my own work. Any contribution from any other party, and any use of generative artificial intelligence technologies have been duly cited and acknowledged;
- is not substantially the same as any other that I have submitted, and;
- is not being concurrently submitted for a degree, diploma or other qualification at the University of Luxembourg or any other University or similar institution except as specified in the text.

With my approval I furthermore confirm the following:

- I have adhered to the rules set out in the University of Luxembourg's Code of Conduct and the Doctoral Education Agreement (DEA)¹, in particular with regard to Research Integrity.
- I have documented all methods, data, and processes truthfully and fully.
- I have mentioned all the significant contributors to the work.
- I am aware that the work may be screened electronically for originality.

I acknowledge that if any issues are raised regarding good research practices based on the review of the thesis, the examination may be postponed pending the outcome of any investigation of such issues. If a degree was conferred, any such subsequently discovered issues may result in the cancellation of the degree.

Approved on 2026-02-13

¹ If applicable (DEA is compulsory since August 2020)

No path in life is ever wasted; every step counts. I would like to express my sincere gratitude to everyone who has been part of my PhD journey.

Abstract

The widespread use of social media has transformed how information is created, shared, and consumed, but it has also enabled the rapid diffusion of misinformation with serious consequences for public health, democratic processes, and social trust. Despite extensive research on misinformation, important gaps remain in understanding how users engage with misleading content during health crises, how fact-checking organizations select claims for verification, and how effective emerging fact-checking interventions are in real-world settings. This thesis addresses these gaps by examining misinformation diffusion and fact-checking practices on social media, with a particular focus on the role and effectiveness of crowdsourced fact-checking approach, *i.e.*, the Community Notes system deployed on X (formerly Twitter).

First, the thesis advances the understanding of user engagement with misinformation during health crises. Using the COVID-19 pandemic as a case study, it shows that misinformation often spans multiple topics and incorporates conspiracy narratives, which increases its attractiveness to users. Crucially, the thesis distinguishes between different user types on social media – those who share external news items and those who react to them on X – and demonstrates that these groups exhibit different engagement patterns. While news sharers contribute to the persistence of misinformation over time, post viewers are more likely to engage with content characterized by topic diversity and conspiratorial framing.

Second, the thesis examines traditional fact-checking practices by investigating how professional fact-checking organizations select politicized statements for verification. Analyzing fact-checked claims mentioning U.S. political elites over a long time span, it uncovers systematic differences between fact-checked true and false statements, partisan asymmetries, and temporal patterns around election periods. These findings shed light on how politicization and emotional framing could influence fact-checking priorities, contributing to debates about neutrality and trust in professional fact-checking.

Third, the thesis provides a comprehensive evaluation of community-based fact-checking through a series of large-scale empirical studies of X's Community Notes system. It offers causal evidence that community notes can substantially (61.2%) reduce the spread of misleading posts and increase the likelihood (94.3%) that authors delete their misleading posts once notes are displayed. However, it also shows that these effects are limited at the platform level because notes often appear too late to intervene during the early and most viral stages of diffusion. In addition, the thesis demonstrates that community fact-checks elicit strong emotional reactions, including increased negativity and moral outrage, highlighting both their corrective power and their role as signals of social norm violations.

Finally, the thesis investigates the algorithm resilience and participation expansion of Community Notes system, exploring how community-based fact-checking systems can be improved. It examines the stability of the consensus-based algorithm used to surface helpful notes and reveals vulnerabilities to post-display polarization and strategic rating behavior. It also evaluates the impact of expanding participation through user-initiated requests for fact-checks, identifying both the promise of scalable, user-led triage and the challenges arising from partisan misalignment between requestors and contributors.

Overall, this thesis makes important contributions to misinformation research by (i) advancing the understanding of misinformation diffusion, (ii) identifying partisan asymmetries in traditional fact-checking practices, (iii) evaluating the effectiveness of community-based fact-checking through a comprehensive and multi-dimensional perspective, and (iv) examining the algorithm resilience and participation expansion of community-based fact-checking. Together, these findings provide actionable insights for social media platforms to develop more reliable, timely, and resilient fact-checking systems.

Contents

Abstract	vii
Contents	ix
1 Introduction	1
1.1 Problem Statement	1
1.2 State of the Art and Research Gaps	2
1.2.1 User Engagement With Misinformation	2
1.2.2 Misinformation Interventions	3
1.3 Research Objectives	5
1.4 Overview of Thesis Structure	6
1.5 Summary of Main Contributions	9
1.6 List of Publications	10
 PART I: MISINFORMATION DIFFUSION AND TRADITIONAL FACT-CHECKING	 13
2 User Engagement With Misinformation During Health Crises	15
2.1 Introduction	15
2.2 Background and Related Work	18
2.2.1 Misinformation on Social Media	18
2.2.2 Misinformation During Public Events/Crises	18
2.2.3 Engagement With COVID-19 Misinformation on Social Media	19
2.2.4 Multi-Topic Context and Conspiracy Theories During COVID-19	20
2.3 Data and Methods	21
2.3.1 Dataset: MM-COVID	21
2.3.2 COVID-19 Topic Spectrum	21
2.3.3 Topic Diversity	23
2.3.4 COVID-19 Conspiracy Theories	24
2.3.5 User Engagement on $\mathbb{X}$	25
2.3.6 Empirical Models	25
2.4 Empirical Results	28
2.4.1 Topic Diversity (RQ1.1 & RQ2.1)	28
2.4.2 News Source Sharer Engagement: News Lifetime and Post Count (RQ1.2 & RQ2.2)	30
2.4.3 Post Viewer Engagement: Reposts, Likes, and Replies (RQ1.2 & RQ2.2)	31
2.4.4 Robustness Checks	33
2.4.5 Summary of Main Findings	34
2.5 Discussion	34
2.5.1 Relevance	34
2.5.2 Research Implications	35
2.5.3 Practical Implications	37
2.5.4 Limitations and Further Research	37
2.6 Conclusion	38
 3 Fact-Checking Selection in Third-Party Organizations	 41
3.1 Introduction	41
3.2 Background	43
3.3 Research Questions	43
3.4 Data and Methods	45
3.4.1 Data Source: Fact-Checked Statements	45

3.4.2	Politicization	45
3.4.3	Other-Condemning Emotions	47
3.4.4	Explanatory Regression Models	47
3.5	Empirical Analysis	48
3.5.1	Mentions of Political Elites (RQ1 & RQ2)	49
3.5.2	Elections (RQ3)	51
3.5.3	Other-Condemning Emotions (RQ4)	52
3.5.4	Exploratory Analysis & Robustness Checks	53
3.6	Discussion	55
3.7	Conclusion	56

PART II: EFFECTIVENESS OF COMMUNITY-BASED FACT-CHECKING 57

4	Overall Engagement Reduction	59
4.1	Introduction	59
4.2	Background and Related Work	62
4.2.1	Misinformation Interventions	62
4.2.2	Community Notes on X/Twitter	63
4.3	Materials and Methods	64
4.3.1	Data Collection	64
4.3.2	Variable Extraction	65
4.3.3	Empirical Models	67
4.4	Empirical Analysis	69
4.4.1	Fact-Checking via Community Notes	69
4.4.2	Analysis at the Feature Roll-Out Level (DiD)	71
4.4.3	Placebo Analysis	74
4.4.4	Analysis at the Note Display Level (RDD)	78
4.4.5	Confounding Analyses	80
4.5	Discussion	84
4.5.1	Research Implications	85
4.5.2	Practical Implications	87
4.5.3	Limitations and Future Research	87
4.6	Conclusion	88
5	Effects of Visibility on Overall Engagement	89
5.1	Introduction	89
5.2	Related Work	91
5.2.1	Social Media Affordances	91
5.2.2	Content Moderation and Self-Censorship	92
5.3	Study 1: Observational Analysis	93
5.3.1	Methods	93
5.3.2	Empirical Results	95
5.4	Study 2: User Study	98
5.4.1	Experiment Setup	98
5.4.2	Experimental Results	98
5.5	Discussion	101
6	Time-Series Repost Reduction	105
6.1	Introduction	105
6.2	Results	107
6.2.1	Difference-in-Differences Estimation	107
6.2.2	Effect on the Cumulative Repost Count	109
6.2.3	Analysis of Reposting Mechanisms	111
6.2.4	Effect on Post Deletion and Suspension/Protection	112

6.3	Discussion	114
6.4	Methods	116
6.4.1	Data Collection	116
6.4.2	Empirical Analysis	116
7	Moral Outrage in Replies	119
7.1	Introduction	119
7.2	Background and Related Work	122
7.2.1	Misinformation on Social Media	122
7.2.2	Content Moderation for Misinformation	123
7.2.3	Sentiments, Basic Emotions, and Moral Outrage on Social Media	123
7.3	Data and Methods	125
7.3.1	Datasets	125
7.3.2	Sentiments and Basic Emotions	126
7.3.3	Political vs. Non-Political Posts	127
7.3.4	Empirical Models	128
7.4	Results	131
7.4.1	Data Overview	131
7.4.2	Analysis of Sentiments and Basic Emotions (RQ1.1)	134
7.4.3	Analysis Across Politic and Non-Political Misleading Posts (RQ1.2)	138
7.4.4	Analysis of Moral Outrage (RQ2.1 & RQ2.2)	141
7.4.5	Robustness Checks	143
7.4.6	Summary of Main Findings	144
7.5	Discussion	145
7.5.1	Relevance	145
7.5.2	Interpretations	145
7.5.3	Implications for Further Design of Community-Based Fact-Checking	147
7.5.4	Limitations and Future Research	148
7.6	Conclusion	149
PART III: RESILIENCE AND PARTICIPATION IN COMMUNITY-BASED FACT-CHECKING		151
8	Consensus Stability	153
8.1	Introduction	153
8.2	Background and Related Work	155
8.3	Data and Methods	157
8.3.1	Dataset	157
8.3.2	Reproduction of Note Selection Algorithm	157
8.3.3	Matrix Factorization for Rater Similarity	158
8.3.4	Estimation Models	159
8.4	Results	160
8.4.1	Note Disappearance (RQ1 & RQ2)	160
8.4.2	Changes in Rating Count and Leaning (RQ3)	161
8.4.3	Similar vs. Dissimilar Raters (RQ4)	162
8.4.4	Counterfactual Analysis	164
8.5	Discussion	165
9	Participation Expansion	167
9.1	Introduction	167
9.2	Background and Related Work	169
9.3	Data and Methods	170
9.3.1	Data Collection	170
9.3.2	Characteristics of Post Content	171
9.3.3	Characteristics of Post Authors	173

9.3.4	External Domain References in Community Notes	174
9.3.5	Identification of Top Writers	174
9.3.6	Replication of Note Selection Algorithm	175
9.3.7	Estimation Models	175
9.4	Empirical Results	176
9.4.1	Likelihood of Receiving Community Notes	176
9.4.2	Request Timing and Request-Fostered Notes	178
9.4.3	Algorithmic Evaluation of Community Notes	180
9.4.4	Summary of Main Findings	185
9.5	Discussion	185
9.5.1	Research Implications	185
9.5.2	Practical Implications	187
9.5.3	Limitations and Future Work	188
10	Conclusion	191
	SUPPLEMENTARY MATERIALS	193
S1	Ethical Considerations	195
S2	Supplementary Materials for Chapter 2	197
S2.1	Estimation Results for Main Analysis	197
S2.2	Cross-Correlations & Variance Inflation Factors	200
S2.3	Robustness Check With Original 26 Topics	201
S2.4	Robustness Check at Theme Level	203
S2.5	Robustness Check With Propensity Score Matching	205
S2.6	Robustness Check With Inclusion of QAnon Conspiracy Theory	207
S2.7	Robustness Check With Exclusion of Potential Debunking Posts	209
S2.8	Robustness Check With BERTopic	211
S2.9	Additional Analysis on Quote Count	213
S3	Supplementary Materials for Chapter 3	215
S3.1	Dataset Overview	215
S3.2	Manual Validation	215
S3.3	Analysis Across Topics	216
S3.4	Analysis Across Incumbent U.S. Presidents	218
S3.5	Political Leanings	220
S3.6	Robustness Checks	221
S3.6.1	Cross-Correlations & Variance Inflation Factors	221
S3.6.2	Statements Without Mixed Veracity	221
S3.6.3	Statements With Manual Validation	222
S3.6.4	Statements Without Mixed Political Elites	222
S3.6.5	Organization-Specific Fixed Effects	222
S3.6.6	Author-Specific Fixed Effects	222
S3.6.7	Propensity Score Matching	223
S4	Supplementary Materials for Chapter 4	231
S5	Supplementary Materials for Chapter 5	247
S5.1	Influencer Bias Chart	247
S5.2	DiD Specification	248
S5.3	HonestDiD Estimation	248
S5.4	Reweighting	249

S6 Supplementary Materials for Chapter 6	251
S6.1 Data Overview	251
S6.2 Descriptive Statistics	253
S6.2.1 Fact-Checking Activity Over Time	253
S6.2.2 Stability of Note Status	254
S6.3 Identification of Topics	255
S6.4 Propensity Score Matching	256
S6.5 Estimation Results	258
S6.5.1 Two-Period ATTs	258
S6.5.2 Parallel Test and Multi-Period ATTs	258
S6.6 Placebo Analyses	264
S6.7 Sensitivity Analyses	268
S6.7.1 Sensitivity Across Response Time	268
S6.7.2 Sensitivity Across Months From Roll-Out	268
S6.7.3 Sensitivity Across Rating Thresholds and Note Helpfulness Scores	269
S6.7.4 Sensitivity Across Poster and Post Characteristics	269
S6.8 Analysis of Reposting Mechanisms	277
S6.8.1 Collection of Reposter Information	277
S6.8.2 Reposter Dataset	278
S6.8.3 Parallel Trend and Treatment Effect	278
S6.8.4 ATTs Across Reposter Characteristics	279
S6.8.5 Sensitivity Across Poster and Reposter Characteristics	279
S6.9 Robustness Checks	293
S6.9.1 ATT Estimation With Post-Level Fixed Effects	293
S6.9.2 ATT Estimation With Zero-Inflated Negative Binomial Regression	293
S6.9.3 Analysis With Restrictions of Note Scores and Pre-Display Engagement	293
S6.9.4 Comparison to Previous Work Analyzing Aggregated Repost Counts	295
S6.9.5 Alternative Before-Display Periods	297
S6.9.6 Analysis With HonestDiD	297
S6.10 Analysis of Deleted and Suspended/Protected Posts	301
S6.10.1 Analysis of Deleted Posts	301
S6.10.2 Analysis of Protected/Suspended Posts	302
S6.11 Changes in X's Algorithms and Policies on Misinformation	302
S6.11.1 Timeline of Changes	303
S6.11.2 Accounting for Algorithmic Effects in Our Study	304
S7 Supplementary Materials for Chapter 7	305
S7.1 Data Overview	305
S7.1.1 Correlations Between Sentiments/Emotions in Source Posts and Sentiments/Emotions in Replies	305
S7.1.2 Sentiments and Emotions in Replies Over Time	306
S7.1.3 Correlations Between Sentiments and Emotions	309
S7.1.4 Correlations Among Emotions	309
S7.2 Descriptive Analysis	310
S7.3 Estimation Results for Main Analysis	311
S7.4 RDD Estimates Across Different Bandwidths	313
S7.5 Sensitivity Analysis	317
S7.6 Moral Outrage Analysis	320
S7.7 Robustness Checks	321
S7.7.1 Variance Inflation Factors	321
S7.7.2 Test for Potential Autocorrelation	322
S7.7.3 Shorter Bandwidth and Placebo Test	322
S7.7.4 Analysis Without Limiting the Number of Replies	323
S7.7.5 Month-Year Fixed Effects	323

S7.7.6	Alternative Lexicon-Based Sentiments and Emotions	324
S7.7.7	Expansion With Sentiments/Emotions in Community Notes	324
S7.7.8	Expansion With Helpfulness Scores of Community Notes	325
S7.7.9	Robustness to the Development of Community Notes Program	325
S8	Supplementary Materials for Chapter 8	339
S8.1	Status Updates on Community Notes	339
S8.2	Reproduction of Note Selection Algorithm	339
S8.3	Matrix Factorization for Rater Similarity	339
S8.4	Characteristics of Posts and Post Authors	340
S8.5	Sensitivity Analysis	341
S8.6	Evaluation of Domain Bias and Quality	341
S9	Supplementary Materials for Chapter 9	343
S9.1	Misleading Criteria	343
S9.2	GPT Prompt	344
S9.3	GPT Validation with PolitiFact	345
S9.4	Likelihood of Receiving Community Notes for Political Posts	346
S9.5	Robustness Check for Top Writers' Post Selection	347
S9.6	Robustness Check for Likelihood of Receiving Community Notes	348
	Bibliography	349

List of Figures

2.1	Topic co-occurrence in true and false news claims	23
2.2	Engagement on $\mathbb{X}$ at news source sharer level and post viewer level	25
2.3	The topic diversity from news claims to news posts	29
2.4	The engagement at news source sharer level	30
2.5	The engagement at post viewer level	32
3.1	The five most frequently mentioned politicians in fact-checked statements	46
3.2	Dataset overview	49
3.3	The politicization ratios of the nine topics that have statistically significant differences during Obama's and Trump's incumbencies	50
3.4	Predicted marginal means for the probabilities of mentions of political elites of either political party, Republicans, and Democrats in fact-checked true and false statements	51
4.1	Daily counts of created notes and fact-checked tweets, smoothed by a 7-day moving average .	70
4.2	Complementary cumulative distribution functions (CCDFs) of the delays	71
4.3	Two-week rolling averages of retweet count and like count in treatment (F-CRH) and control (F-NCRH) groups from 2022-07 to 2023-04	75
4.4	Two-week rolling averages of retweet count and like count in treatment (F-CRH) and control (T-NMR) groups from 2022-07 to 2023-04	77
4.5	The averages of retweet count and like count across the note score points	79
5.1	Research overview	90
5.2	Analysis of weekly effects of making likes private on the number of likes	96
5.3	The estimated aggregated effects across account categories in Study 1	97
5.4	Reported channels through which participants first became aware of $\mathbb{X}$'s change to private like visibility	99
5.5	Participants' self-rated likelihood of liking posts on high-reputational-risk and low-reputational-risk topics if their likes were public or private ($N = 203$)	100
6.1	Community notes reduce the spread of misleading posts on $\mathbb{X}$	109
6.2	Sensitivity analysis	110
6.3	Effect of community notes on the cumulative repost count of misleading posts on $\mathbb{X}$	111
6.4	Analysis of reposting mechanisms	112
6.5	Effect of community notes on post deletion and suspension/protection	113
7.1	Research overview	126
7.2	Summary statistics for misleading source posts, direct replies, and display of community notes	132
7.3	The estimated coefficients for the independent variables	135
7.4	The predicted effects of the display of community notes on sentiments and emotions in replies	137
7.5	The RDD coefficient estimates for sentiments and emotions in replies across three different bandwidths	138
7.6	The estimated coefficients for the independent variables	139
7.7	Moral outrage in replies to misleading posts increases after the display of community notes . .	141
8.1	Workflow of the note selection algorithm	154
8.2	Estimation results for the likelihood of note disappearance	161
8.3	Changes in rating count and rating leaning after the initial display of community notes	162
8.4	Discrepancies in rating leaning between similar and dissimilar raters	162
8.5	Evaluation of polarized ratings and counterfactual analysis of the note selection algorithm . .	163

9.1	The introduction of “Request Community Note” feature on $\mathbb{X}$	168
9.2	The estimation results for the logistic regression model predicting the likelihood that a requested post receives a community note	177
9.3	The statistics for the request timing relative note writing	178
9.4	The estimation results for the logistic regression model predicting the likelihood that a requested post receives a top-writer note after the request threshold	179
9.5	Overview of note evaluations and source domains in community notes	181
9.6	The estimation results for the linear regression model predicting note helpfulness	182
9.7	The estimation results for the linear regression model predicting note polarization	184
S3.1	Topics in fact-checked statements	217
S4.1	The illustrations of F-CRH (treatment) and F-NCRH (control) groups	232
S4.2	Parallel tests of retweet count and like count between the treatment and control groups	233
S4.3	CCDFs of retweet count and like count based on the fact-checked tweets from matched F-CRH (Misleading) and T-NMR (Not Misleading) groups	234
S4.4	CCDFs of retweet count and like count based on the fact-checked tweets from matched F-CRH (Misleading) and T-NMR (Not Misleading) groups, during different time periods	235
S5.1	Influencer bias chart	247
S5.2	The HonestDiD estimates for the post-change weekly periods	249
S5.3	Demographics of participants	249
S6.1	Data overview	252
S6.2	The ratio of displayed notes relative to all the community notes over time	254
S6.3	The distributions of propensity score before and after one-to-one propensity score matching	257
S6.4	Equivalence test for parallel trend	260
S6.5	The original DiD coefficient estimates and robust confidence intervals under different M_{bar}	263
S6.6	The hourly DiD coefficient estimates	264
S6.7	Placebo analyses for the effectiveness of community notes	265
S6.8	The estimated ATTs of community notes depending on the number of ratings	270
S6.9	The distributions of the propensity score for collected posts in the treatment and control groups	277
S6.10	The estimated ATTs across different groups of reposters based on posts separated by poster characteristics	279
S6.11	The estimated ATTs across different groups of reposters based on posts separated by poster characteristics	280
S6.12	The estimated ATTs across different groups of reposters based on posts separated by post characteristics	281
S6.13	The estimated ATTs across different groups of reposters based on posts separated by post characteristics	282
S6.14	Analysis with restrictions of note scores and pre-display engagement	296
S6.15	CCDFs showing the actually observed repost count for source posts with displayed community notes and the predicted repost count that the source posts would have received in the absence of community notes display	297
S6.16	The original DiD coefficient estimate for the aggregated ATT after the display of community notes and its robust confidence intervals under different M_{bar}	299
S7.1	The Pearson correlations between sentiments (emotions) in replies and source sentiments (emotions) in misleading posts	305
S7.2	The hourly averages in sentiments and emotions across hours from note display	308
S7.3	The sentiments and emotions in replies before and after the display of community notes	310
S7.4	The lagged correlation estimates across sentiments and emotions in replies in Chronological, Shuffled, and Reversed orders, respectively	322

S8.1	The distribution of estimated Core note intercepts (helpfulness) and estimated Core note factors (polarization) from the note selection algorithm	339
S8.2	Analysis of MF-based model for rater representations	340
S8.3	Sensitivity analysis	341
S9.1	The note-writing interface for Community Notes contributors	343
S9.2	The estimation results for the logistic regression model predicting the likelihood that a requested political post receives a community note	346
S9.3	The estimation results for the logistic regression model predicting the likelihood that a requested post receives a top-writer note after the request threshold	347
S9.4	The estimation results for the logistic regression model predicting the likelihood that a requested post receives a community note	348

List of Tables

1.1	Thesis structure	7
2.1	COVID-19 topic spectrum	22
2.2	The examples of topic seeds (bioweapon, vaccine, election, and quarantine) and their 10 most similar words	22
2.3	COVID-19 conspiracy theories and corresponding keywords (phrases)	24
2.4	Dataset overview	27
3.1	Examples of fact-checked true and false statements	45
3.2	Estimation results for mixed logistic regression models predicting whether the fact-checked statements mention political elites of either political party, Republicans, or Democrats	50
3.3	Estimation results for mixed linear regression models predicting other-condemning emotions in the fact-checked statements	53
4.1	Regression results for response number and time	72
4.2	DiD regression results for retweet count and like count within F-CRH (treatment) and F-NCRH (control) groups	74
4.3	DiD regression results for retweet count and like count within F-CRH (treatment) and T-NMR (control) groups	76
4.4	RDD regression results for retweet count and like count within F-CRH (treatment) and F-NCRH (control) groups	80
4.5	Regression results for retweet rate and like rate within F-CRH (treatment) and OCN (control) groups	84
5.1	Overview of accounts, posts, and their associated engagement of likes and reposts	94
7.1	Dataset overview	127
7.2	<i>SUEST</i> tests for the equality of common coefficients across RDD models with different bandwidths	138
8.1	Dataset overview	157
9.1	Overview of requested posts	171
S2.1	Estimation results for the topic diversity in news claims, the topic diversity in news posts, and the changes in topic diversity from news claims to associated posts	197
S2.2	Estimation results for news lifetime and post count	198
S2.3	Estimation results for repost count, like count, and reply count	198

S2.4	Estimation results for repost count, like count, and reply count in the subset of false news posts	199
S2.5	Correlations of variables in news claims	200
S2.6	Correlations of variables in news posts	200
S2.7	VIFs	200
S2.8	Estimation results for the topic diversity in news claims, the topic diversity in news posts, and the changes in topic diversity from news claims to associated posts	201
S2.9	Estimation results for news lifetime and post count	202
S2.10	Estimation results for repost count, like count, and reply count	202
S2.11	Estimation results for the topic diversity in news claims, the topic diversity in news posts, and the changes in topic diversity from news claims to associated posts	203
S2.12	Estimation results for news lifetime and post count	204
S2.13	Estimation results for repost count, like count, and reply count	204
S2.14	Estimation results for the topic diversity in news claims, the topic diversity in news posts, and the changes in topic diversity from news claims to associated posts	205
S2.15	Estimation results for news lifetime and post count	206
S2.16	Estimation results for repost count, like count, and reply count	206
S2.17	Estimation results for the topic diversity in news claims, the topic diversity in news posts, and the changes in topic diversity from news claims to associated posts	207
S2.18	Estimation results for news lifetime and post count	208
S2.19	Estimation results for repost count, like count, and reply count	208
S2.20	Estimation results for the topic diversity in news claims, the topic diversity in news posts, and the changes in topic diversity from news claims to associated posts	209
S2.21	Estimation results for news lifetime and post count	210
S2.22	Estimation results for repost count, like count, and reply count	210
S2.23	Estimation results for the topic diversity in news claims, the topic diversity in news posts, and the changes in topic diversity from news claims to associated posts	211
S2.24	Estimation results for news lifetime and post count	212
S2.25	Estimation results for repost count, like count, and reply count	212
S2.26	Estimation results for quote count	213
S3.1	Dataset overview	215
S3.2	The twenty most frequent topic tags attached to fact-checked statements	216
S3.3	Estimation results for mixed logistic regression models predicting whether the fact-checked statements mention political elites of either political party, Republicans, or Democrats	218
S3.4	Estimation results for mixed logistic regression models predicting whether the fact-checked statements mention political elites of either political party, Republicans, or Democrats	219
S3.5	Estimation results for mixed linear regression models predicting other-condemning emotions in the fact-checked statements	219
S3.6	Estimation results for mixed linear regression models predicting partisan leanings in the fact-checked statements	220
S3.7	Cross-correlations among independent variables	221
S3.8	Variance Inflation Factors (VIFs) for the independent variables	221
S3.9	Estimation results for mixed logistic regression models predicting whether the fact-checked statements mention political elites of either political party, Republicans, or Democrats	222
S3.10	Estimation results for mixed linear regression models predicting other-condemning emotions in the fact-checked statements	223
S3.11	Estimation results for mixed logistic regression models predicting whether the fact-checked statements mention political elites of either political party, Republicans, or Democrats	224
S3.12	Estimation results for mixed linear regression models predicting other-condemning emotions in the fact-checked statements	224
S3.13	Estimation results for mixed logistic regression models predicting whether the fact-checked statements mention political elites of either political party, Republicans, or Democrats	225
S3.14	Estimation results for mixed linear regression models predicting other-condemning emotions in the fact-checked statements	225

S3.15	Estimation results for mixed logistic regression models predicting whether the fact-checked statements mention political elites of either political party, Republicans, or Democrats	226
S3.16	Estimation results for mixed linear regression models predicting other-condemning emotions in the fact-checked statements	226
S3.17	Estimation results for fixed logistic regression models predicting whether the fact-checked statements mention political elites of either political party, Republicans, or Democrats	227
S3.18	Estimation results for fixed linear regression models predicting other-condemning emotions in the fact-checked statements	227
S3.19	Estimation results for mixed logistic regression models predicting whether the fact-checked statements mention political elites of either political party, Republicans, or Democrats	228
S3.20	Estimation results for mixed linear regression models predicting other-condemning emotions in the fact-checked statements	228
S3.21	Estimation results for mixed logistic regression models predicting whether the fact-checked statements mention political elites of either political party, Republicans, or Democrats	229
S3.22	Estimation results for mixed linear regression models predicting other-condemning emotions in the fact-checked statements	229
S4.1	Topic list with corresponding domains	236
S4.2	Descriptive statistics	236
S4.3	DiD regression results for retweet count and like count within F-CRH (treatment) and F-NCRH (control) groups between Pilot and Roll-Out periods	237
S4.4	DiD regression results for retweet count and like count within T-NMR (treatment) and F-NCRH (control) groups	238
S4.5	DiD regression results for retweet count and like count within F-CRH (treatment) and T-NMR (control) groups under the limitation of the top 25% delay in note display	239
S4.6	DiD regression results for retweet count and like count between Pilot and Roll-Out periods in placebo tests	240
S4.7	DiD regression results with external controls for retweet count and like count within F-CRH (treatment) and F-NCRH (control) groups	241
S4.8	DiD regression results with external controls for retweet count plus, like count plus, and retweets including quotes within F-CRH (treatment) and F-NCRH (control) groups	242
S4.9	DiD regression results with external controls for retweet count and like count within F-CRH (treatment) and F-NCRH (control) matched groups during different periods	243
S4.10	DiD regression results with external controls for retweet count and like count within F-CRH (treatment) and OCN (control) groups	244
S4.11	Regression results for retweet count parallel test between F-CRH and F-CNRH groups	245
S4.12	Regression results for like count parallel test between F-CRH and F-CNRH groups	245
S4.13	Regression results for retweet count parallel test between T-NMR and F-CNRH groups	245
S4.14	Regression results for like count parallel test between T-NMR and F-CNRH groups	246
S4.15	Regression results for retweet count parallel test between F-CRH and T-NMR groups	246
S4.16	Regression results for like count parallel test between F-CRH and T-NMR groups	246
S6.1	Dataset overview	254
S6.2	Out-of-sample performance in topic prediction	256
S6.3	Regression results for the evaluation of propensity score matching	257
S6.4	Regression results for two-period ATT estimation	260
S6.5	Regression results for parallel test and multi-period ATTs estimation	261
S6.6	Regression results parallel trends test with alternative baseline periods	262
S6.7	Regression results for parallel test and multi-period ATTs estimation	264
S6.8	Regression results for the multi-period ATT estimation with the replacement of treatment group	266
S6.9	Regression results for two-period ATT estimation with the replacement of treatment group	267
S6.10	Regression results for two-period ATT estimations over response time	268
S6.11	Regression results for two-period ATT estimations across months from the roll-out of “Community Notes” program	272
S6.12	Regression results for two-period ATT estimations across rating thresholds	273

S6.13	Regression results for two-period ATT estimations depending on the stability of note helpful status	273
S6.14	Regression results for two-period ATT estimations across note helpfulness scores	273
S6.15	Regression results for two-period ATT estimations across user characteristics	274
S6.16	Regression results for two-period ATT estimations across post characteristics	275
S6.17	Regression results for two-period ATT estimations across post characteristics	276
S6.18	Overview of the dataset for reposter analysis	278
S6.19	Regression results for parallel test and multi-period ATTs estimation in the subset of posts for reposter analysis	283
S6.20	Regression results for two-period ATT estimation in the subset of posts for reposter analysis	284
S6.21	Regression results for two-period ATT estimations across reposter characteristics	285
S6.22	Regression results for two-period ATT estimations, based on posts from left-leaning posters and across reposter characteristics	285
S6.23	Regression results for two-period ATT estimations, based on posts from right-leaning posters and across reposter characteristics	285
S6.24	Regression results for two-period ATT estimations, based on posts from high-misinformation-exposure posters and across reposter characteristics	285
S6.25	Regression results for two-period ATT estimations, based on posts from low-misinformation-exposure posters and across reposter characteristics	286
S6.26	Regression results for two-period ATT estimations, based on posts from verified posters and across reposter characteristics	286
S6.27	Regression results for two-period ATT estimations, based on posts from non-verified posters and across reposter characteristics	286
S6.28	Regression results for two-period ATT estimations, based on posts from older accounts and across reposter characteristics	286
S6.29	Regression results for two-period ATT estimations, based on posts from younger accounts and across reposter characteristics	287
S6.30	Regression results for two-period ATT estimations, based on posts from accounts with higher follower counts and across reposter characteristics	287
S6.31	Regression results for two-period ATT estimations, based on posts from accounts with lower follower counts and across reposter characteristics	287
S6.32	Regression results for two-period ATT estimations, based on posts from accounts with higher followee counts and across reposter characteristics	287
S6.33	Regression results for two-period ATT estimations, based on posts from accounts with lower followee counts and across reposter characteristics	288
S6.34	Regression results for two-period ATT estimations, based on longer posts and across reposter characteristics	288
S6.35	Regression results for two-period ATT estimations, based on shorter posts and across reposter characteristics	288
S6.36	Regression results for two-period ATT estimations, based on media posts and across reposter characteristics	288
S6.37	Regression results for two-period ATT estimations, based on non-media posts and across reposter characteristics	289
S6.38	Regression results for two-period ATT estimations, based on high-positive posts and across reposter characteristics	289
S6.39	Regression results for two-period ATT estimations, based on low-positive posts and across reposter characteristics	289
S6.40	Regression results for two-period ATT estimations, based on high-negative posts and across reposter characteristics	289
S6.41	Regression results for two-period ATT estimations, based on low-negative posts and across reposter characteristics	290
S6.42	Regression results for two-period ATT estimations, based on economy-related posts and across reposter characteristics	290

S6.43	Regression results for two-period ATT estimations, based on non-economy-related posts and across reposter characteristics	290
S6.44	Regression results for two-period ATT estimations, based on health-related posts and across reposter characteristics	290
S6.45	Regression results for two-period ATT estimations, based on non-health-related posts and across reposter characteristics	291
S6.46	Regression results for two-period ATT estimations, based on political posts and across reposter characteristics	291
S6.47	Regression results for two-period ATT estimations, based on non-political posts and across reposter characteristics	291
S6.48	Regression results for two-period ATT estimations, based on science-related posts and across reposter characteristics	291
S6.49	Regression results for two-period ATT estimations, based on non-science-related posts and across reposter characteristics	292
S6.50	Regression results for two-period ATT estimation with post-level fixed effects and zero-inflated negative binomial regressions	294
S6.51	Regression results for the treatment effect estimation with the restrictions on note scores and pre-display engagement and the early stage of Community Notes program	298
S6.52	Regression results for two-period ATT estimation with a before-display period of one hour and a before-display period of two hours	298
S6.53	Regression results for multi-period ATTs estimation, treating the entire after-display period as a single interval	300
S6.54	Estimation results for RDD logistic regression models predicting deletion of posts	302
S6.55	Estimation results for RDD logistic regression models predicting suspension/protection of posts	302
S7.1	Estimation results for positive and negative sentiments in replies before note display	306
S7.2	Estimation results for anger, disgust, fear, joy, sadness, and surprise in replies before note display	307
S7.3	Estimation results for positive and negative sentiments in replies	307
S7.4	Estimation results for anger, disgust, fear, joy, sadness, and surprise in replies	308
S7.5	The Pearson correlations between sentiments and emotions with two-sided p-values	309
S7.6	The Pearson correlations among emotions with two-sided p-values	309
S7.7	Estimation results for positive sentiment and negative sentiment in replies within reply lifespan	311
S7.8	Estimation results for anger, disgust, and fear in replies within reply lifespan	312
S7.9	Estimation results for joy, sadness, and surprise in replies within reply lifespan	312
S7.10	Estimation results for positive sentiment and negative sentiment in replies within 16-hour bandwidth	313
S7.11	Estimation results for anger, disgust, fear, joy, sadness, and surprise in replies within 16-hour bandwidth	314
S7.12	Estimation results for positive sentiment and negative sentiment in replies within one-week bandwidth	314
S7.13	Estimation results for anger, disgust, fear, joy, sadness, and surprise in replies within one-week bandwidth	315
S7.14	Estimation results for positive sentiment and negative sentiment in replies within reply lifespan	315
S7.15	Estimation results for anger, disgust, fear, joy, sadness, and surprise in replies within reply lifespan	316
S7.16	Estimation results for positive and negative sentiments in replies to political and non-political misleading source posts	317
S7.17	Estimation results for anger, disgust, and fear in replies to political and non-political misleading source posts	318
S7.18	Estimation results for joy, sadness, and surprise in replies to political and non-political misleading source posts	319
S7.19	Estimation results for moral outrage in replies across all misleading posts, political misleading posts, and non-political misleading posts	320
S7.20	Estimation results for moral outrage based on DOC and other-condemning moral emotions in replies	321

S7.21	VIFs for independent variables	321
S7.22	The lagged coefficient estimates across sentiments and emotions in replies	322
S7.23	The Pearson correlations between source sentiments/emotions in original posts and lagged sentiments/emotions in replies with two-sided p-values	323
S7.24	Estimation results for positive and negative sentiments in replies	323
S7.25	Estimation results for anger, disgust, fear, joy, sadness, surprise, and moral outrage in replies .	324
S7.26	Estimation results for positive sentiment and negative sentiment in replies within 2-hour bandwidth	325
S7.27	Estimation results for anger, disgust, fear, joy, sadness, surprise, and moral outrage in replies within 2-hour bandwidth	326
S7.28	Estimation results for positive sentiment and negative sentiment in replies within 2-hour bandwidth	327
S7.29	Estimation results for anger, disgust, fear, joy, sadness, surprise, and moral outrage in replies within 2-hour bandwidth	327
S7.30	Estimation results for positive and negative sentiments in replies without number limit	328
S7.31	Estimation results for anger, disgust, fear, joy, sadness, surprise, and moral outrage in replies without number limit	328
S7.32	Estimation results for positive sentiment and negative sentiment in replies	329
S7.33	Estimation results for emotions in replies: anger, disgust, fear, joy, sadness, and surprise . . .	329
S7.34	Estimation results for positive sentiment and negative sentiment in replies	330
S7.35	Estimation results for emotions in replies: anger, disgust, fear, joy, sadness, and surprise . . .	330
S7.36	Estimation results for moral outrage across robustness checks	331
S7.37	Estimation results for sentiment polarity and emotions in replies: polarity, anger, disgust, and fear	332
S7.38	Estimation results for emotions in replies: joy, sadness, and surprise	332
S7.39	Estimation results for positive sentiment and negative sentiment in replies	333
S7.40	Estimation results for anger, disgust, fear, joy, sadness, and surprise in replies	334
S7.41	Estimation results for positive and negative sentiments in replies	335
S7.42	Estimation results for anger, disgust, fear, joy, sadness, surprise, and moral outrage in replies .	335
S7.43	Estimation results for positive and negative sentiments in replies	336
S7.44	Estimation results for anger, disgust, fear, joy, sadness, surprise, and moral outrage in replies .	336
S7.45	Estimation results for positive and negative sentiments in replies	337
S7.46	Estimation results for anger, disgust, fear, joy, sadness, surprise, and moral outrage in replies .	337

Nowadays, social media platforms – such as X/Twitter, Facebook, and TikTok – have transformed information exchange, enabling public users to create and engage with online content globally and instantaneously. However, the absence of quality control mechanisms inherent in traditional media has led to a significant rise in misinformation on social media. Particularly, the spread of misinformation on social media has had far-reaching consequences across many critical societal domains ranging from politics to public health, thus becoming a concerning global issue [1–4]. Consequently, reducing user engagement with online misinformation has become a pressing challenge for social media platforms, policymakers, and researchers [5]. In response, a substantial body of research has emerged in recent years to address the definition, characterization, detection, and intervention of online misinformation [1, 6, 7]. Despite these efforts, the problem remains far from being resolved, and social media platforms continue to face increasing pressure to implement effective measures that can reduce the spread of misinformation at scale. Against this backdrop, this thesis seeks to advance the understanding of user engagement with online misinformation, examine fact-checking activities, and evaluate fact-checking effectiveness to inform the design of reliable, scalable interventions for mitigating the spread of misinformation on social media platforms.

1.1 Problem Statement

In this thesis, we define misinformation as false or misleading information that is contradicted by empirical evidence or is inconsistent with common or expert consensus [1, 7, 8].¹ When people believe and give credibility to misinformation, the notion of truth becomes corrupted, leading to potentially harmful consequences. This is particularly problematic in the digital age, where users are exposed to persuasive online information with uncertain origins and a lack of authenticity [12].

In contemporary democratic politics, social media has emerged as the main platform for political propaganda. Dedicated misinformation designed to pursue political agendas easily gets viral engagement online [13–15]. During political elections, people are exposed to online misinformation more often than usual, which could influence individuals' opinions and voting behaviors, and even undermine elections [16–20]. For example, during the 2016 U.S. presidential election, on average, each U.S. adult saw and remembered one or perhaps several false stories with clear partisan leanings on social media [21]. Similar problems also arose in the Global South, for example, during the 2018 Brazilian presidential election [19, 20]. Politics and finance are closely linked, and political false information can accordingly incite shocks in the financial market. For instance, in 2013, \$130 billion in stock value was wiped out in a few minutes following a false tweet indicating that an “explosion” injured Barack Obama [22]. Additionally, false information about companies or economic indicators can also lead to stock market volatility, potentially resulting in financial losses for investors and businesses [23]. Online misinformation can even pose severe risks to public health. For example, the COVID-19 pandemic has

1.1	Problem Statement	1
1.2	State of the Art and Research Gaps	2
1.2.1	User Engagement With Misinformation	2
1.2.2	Misinformation Interventions	3
1.3	Research Objectives . . .	5
1.4	Overview of Thesis Structure	6
1.5	Summary of Main Contributions	9
1.6	List of Publications	10

1: In previous research, a variety of terms have been used to refer to false and misleading information. Examples include “misinformation,” “disinformation,” and “fake news,” each of which places a different emphasis on the underlying characteristics. Disinformation indicates a subset of misinformation that is spread with the intention of harm or deception [7, 8]. Notably, detecting intentions, especially those not explicitly expressed in the content, can be challenging [9]. Fake news is a type of misinformation in the form of news that is often fabricated to mislead people, especially in the political and financial fields [7, 10, 11]. In our work, we use the term “misinformation” as an umbrella as it does not impose assumptions on the motif or the format of the content [1, 7, 8].

been accompanied by an “infodemic” around the world, characterized by the dissemination of a plethora of online misinformation regarding the virus, its origin, and potential treatments or prevention strategies [24]. A global survey demonstrated that a substantial proportion of people view COVID-19 misinformation as reliable [25]. Beliefs in misinformation during COVID-19 influenced people’s hesitancy and refusal of getting vaccinated, which affected their risk of infection [2, 3, 25–28].

While some critics argue that the harms of misinformation may be overstated, with exposure to falsehoods often concentrated among narrow fringe groups [29–31], the negative repercussions of online misinformation on modern societies have been repeatedly observed in recent years, especially in the context of elections and crises [17, 18, 21, 32]. Therefore, the fight against the spread of misinformation on social media is both warranted and required [33].

1.2 State of the Art and Research Gaps

1.2.1 User Engagement With Misinformation

Before being fact-checked, both true and false news items that contain uncertainty are often treated as rumors and can be widely shared among the public [34]. Users tend not to question the credibility of information unless it violates their preconceptions or they are incentivized to do so [5]. The wide spread of rumors, both true and false, is therefore influenced by specific content features, such as topics and emotions, and how users react to these content features [34]. Hence, the driving factors of information diffusion and their use in misinformation need to be comprehensively explored to understand users’ engagement with misinformation.

Prior research shows that misinformation is often associated with socially salient issues or events characterized by high uncertainty and widespread public attention. Such content tends to be more easily shared, can polarize public opinion, and may even trigger social panic [5, 14, 35]. In addition to topical relevance, writing style plays a critical role in attracting attention and persuading users to believe and further disseminate content. For example, misinformation producers often engage in dense linking and mutual reinforcement, which amplifies the spread of false narratives and conspiracy theories [36].

These dynamics become particularly pronounced during health crises, when multiple aspects of society and individual behavior become tightly interconnected. Online discussions during such periods often span multiple domains and exhibit strong cross-topic interactions, contributing to a complex and heterogeneous information environment [37, 38]. Moreover, conspiracy theories frequently emerge during health crises, framing major public events as the result of secret and malicious actions by powerful actors [39–41]. The integration of conspiracy theories within misinformation narratives serves as an effective strategy to deliberately interconnect various topics [37, 42].

RG1: *Despite these insights, online engagement with misinformation in multi-topic information environments during health crises remains insufficiently understood.*²

2: First research gap (RG1)

1.2.2 Misinformation Interventions

Intervention Types and Traditional Fact-Checking

To reduce the spread of misinformation, various interventions have been proposed to ensure that people remain accurately informed. These interventions can be broadly distinguished into two categories, namely, prebunking and debunking [1]. Prebunking is a kind of pre-emptive intervention and seeks to help people recognize and resist subsequently encountered misinformation [1, 7]. Prebunking interventions include factually correct information presentation, pre-emptive correction, digital media literacy training, and generic misinformation warning before exposure to misinformation. Debunking, also known as fact-checking, involves responding to specific instances of misinformation after exposure to demonstrate why they are false [1, 7, 43]. Reliable and trustworthy fact-checking plays a crucial role in promoting accurate information that can inform public opinion and decision-making. Fact-checking organizations such as PolitiFact (politifact.com) and Snopes (snopes.com) are vital for identifying and debunking misinformation. Their fact-checking assessments are supposed to help users to identify misleading statements in order to curb their spread (*e.g.*, on social media). Previous studies suggest that fact-checking can indeed improve misinformation discernment and reduce sharing intentions for misinformation [44]. In practice, however, limited resources often force fact-checkers to prioritize statements for fact-checking. Although their assessments are generally considered accurate [14], the necessity of choosing specific claims to fact-check can invite criticism due to the subjective nature of the decision. Critics argue that the selection of claims can reflect political preferences, potentially skewing results in favor of certain political identities (typically towards the political left). This is problematic as the perception of bias and subjectivity can erode public trust in fact-checking organizations. If people believe that fact-checkers are not impartial, they may be less likely to trust their findings, regardless of the evidence presented.

RG2: *Despite these vocal concerns, however, empirical evidence on how fact-checking organizations pick up politicized misinformation for fact-checking is missing.*³

3: Second research gap (RG2)

Advances in Fact-Checking Approaches

Beyond concerns about selection bias and public trust, the scarcity of expert fact-checkers, coupled with the relentless pace of online content, hampers the timely and scalable implementation of professional fact-checking [45]. To address the limitations of professional fact-checking, crowdsourced (or community-based) fact-checking approaches have been proposed, leveraging the wisdom of crowds to achieve broader and faster coverage of online content [46, 47]. Early explorations of crowdsourced verification demonstrate that journalists and online volunteer communities can collaboratively assess information accuracy and provide real-time event coverage during emergencies [48]. Furthermore, hybrid crowdsourcing processes that involve both local and remote participants help enhance the completeness of information [49]. Although individual assessments may be subject to bias and noise [50], aggregated judgments—even from relatively small crowds—have been shown to be reliable and comparable in accuracy to expert evaluations [46, 51–54]. Moreover, crowdsourced assessments have the potential to mitigate public trust issues associated with expert fact-checks [55, 56]. Together,

these studies provide important foundations for the development of community-based fact-checking systems.

Meanwhile, automatic fact-checking powered by machine learning models continues to develop, offering increasingly sophisticated tools for identifying misinformation. Traditional detection models are typically trained on surface-level content or contextual features such as linguistic patterns, semantic similarity, propagation structures, or user metadata [57]. Although these approaches have proven useful in identifying repeated misinformation or content with distinctive stylistic markers, particularly in constrained domains or benchmark datasets, they often struggle with more subtle, novel, or context-dependent cases. In particular, such models lack the ability to reason over evidence, integrate external knowledge sources, or adapt quickly to evolving narratives. As a result, their performance in real-world fact-checking scenarios often falls short, especially when confronted with emerging claims that deviate from the data on which they were trained [58].

Recent advances in Large Language Models (LLMs), such as OpenAI’s GPT series, introduce novel opportunities to overcome the shortcomings of traditional machine learning models in fact-checking. LLMs exhibit strong contextual understanding and can simulate human-like reasoning processes, allowing them to evaluate claims not only on textual cues but also in relation to broader world knowledge. They can flexibly incorporate external evidence through retrieval-augmented generation, adapt to diverse domains with minimal task-specific training, and generate explanations that improve transparency in the fact-checking process [57, 59]. Previous work suggests that LLMs can achieve human-comparable performance in misinformation detection tasks [60–64]. For instance, ChatGPT has been shown to achieve an accuracy of approximately 90% in identifying false headlines, demonstrating the potential of LLMs to support large-scale misinformation detection [61]. At the same time, LLM-based fact-checking is not without challenges. LLMs tend to adopt a more structured evaluation strategy, whereas human annotators often display greater variability in their use of evaluative criteria, particularly for borderline or ambiguous claims [62]. More critically, their influence on user perception can be harmful: when LLMs mislabel true claims as false, they risk reducing belief in accurate information; conversely, when they express uncertainty about false claims, they may inadvertently increase belief in misinformation [61, 65]. These limitations underscore that, while LLMs can perform well in detecting misinformation, human judgment remains essential for deciding what content to surface to users, helping to mitigate unintended harms.

Crowdsourced Fact-Checking in Practice: X’s Community Notes Program

Motivated by the necessity of human-centered approaches and the demonstrated potential of crowdsourced fact-checking, X introduced Community Notes, a platform-wide and community-based fact-checking system that engages a diverse set of users, referred to as *contributors*, in collaboratively annotating potentially misleading content at scale [66]. Specifically, Community Notes contributors can append contextual information or corrections to potentially misleading posts, offering alternative perspectives or clarifications. Contributors are required to cite external sources in their community notes, and these notes are then continuously evaluated through peer ratings from other contributors. The statuses of

notes are determined by an algorithm that calculates the helpfulness score based on the ratings made by the contributors from different perspectives [67]. Only notes that mark posts as misleading and receive high helpfulness scores earn the status of “Helpful” and are displayed on the corresponding tweets.

The deployment of Community Notes on X represents the first large-scale, real-world implementation of a crowdsourced, community-based fact-checking system on a major social media platform, offering a promising content moderation approach that complements traditional expert-based fact-checking in combating misinformation.

RG3: *Yet, empirical evidence on the effectiveness of community-based fact-checking in reducing engagement with misinformation on social media platforms remains missing.*⁴

4: Third research gap (RG3)

Unlike expert fact-checkers, who provide rigorous and authoritative assessments, community contributors vary widely in expertise, consistency, and impartiality across both note-writing and rating processes [50, 54, 68, 69]. Consequently, aggregation mechanisms that rely on majority agreement to surface community notes can render community-based systems vulnerable to manipulation or one-sided ratings from ideologically homogeneous groups. Research shows that politically balanced groups produce more reliable and accurate evaluations than homogeneous ones [46, 47]. Therefore, X develops a bridging algorithm to feature objectively informative community notes that are subjectively perceived to be helpful across heterogeneous contributors [67]. Notably, the algorithm updates note statutes regularly (e.g., hourly) by incorporating new ratings.

RG4: *While this note selection algorithm can effectively feature helpful notes that achieve cross-group consensus and are factually correct [70], the stability of such consensus after notes are displayed remains underexplored.*⁵

5: Fourth research gap (RG4)

Understanding this stability is important for informing the design of more robust and resilient Community Notes algorithms.

Beyond community contributors, prior research has also explored on-demand fact-checking models that enable ordinary users to request fact-checks for content they wish to have checked [71]. In line with this idea, all users on X can participate indirectly through the recently launched Request Community Note function, which allows any account with a verified phone number to submit requests for posts they believe would benefit from a community note. When enough requests are submitted, contributors can see alerts highlighting these posts and can choose to write a note in response. This division of roles – between note writers, note raters, and requestors – illustrates how Community Notes system distributes responsibilities across different layers of the user base, thereby expanding participation beyond the core contributor group.

RG5: *However, despite its potential significance, this new function has not yet been systematically studied, leaving open questions about how it shapes the Community Notes system—for instance, its effects on post selection, note generation, and the quality of resulting notes.*⁶

6: Fifth research gap (RG5)

1.3 Research Objectives

Based on the identified research gaps, this thesis pursues four research objectives to advance the understanding of misinformation diffusion on social media, examine traditional fact-checking activities, and evaluate an innovative fact-checking approach, namely, Community Notes, to

inform the design of reliable and scalable interventions for mitigating the spread of misinformation on social media platforms.

- RO1** To examine how users engage with misinformation during health crises, focusing on topic diversity and conspiracy theories, and to compare engagement patterns between news source sharers and post viewers. This objective addresses **RG1**.
- RO2** To investigate how professional fact-checking platforms select and prioritize politicized misinformation for verification. This objective addresses **RG2**.
- RO3** To evaluate the effectiveness of community-based fact-checking in reducing engagement with misinformation, taking into account overall engagement patterns, confounding platform design changes, temporal dynamics of engagement, and users' emotional responses. This objective addresses **RG3**.
- RO4** To assess the optimization and evolution of community-based fact-checking systems, with a focus on the resilience of consensus-based algorithms against manipulation and the system-level effects of expanded user participation. This objective addresses **RG4** and **RG5**.

1.4 Overview of Thesis Structure

This thesis is organized into three main research parts across eight core chapters to address its research objectives. Table 1.1 summarizes how the chapters correspond to the research objectives and the publications on which they are based.

Part I: Misinformation diffusion and traditional fact-checking. Part I includes two chapters examining user engagement with misinformation during health crises and the selection of politicized statements by professional fact-checkers.

Chapter 2 addresses **RO1** by exploring how people engage with misinformation during health crises, using the COVID-19 pandemic as an example. It shows that misleading content often combines multiple topics and conspiracy narratives, making it more attractive to users. Importantly, this chapter distinguishes between those who share news items (*i.e.*, news source sharers) and those who react to them (*i.e.*, post viewers), revealing that these two groups behave differently. While news sharers help misinformation persist over time, viewers are more likely to engage with misleading content that spans diverse topics or includes conspiracy theories. These findings suggest that effective interventions need to limit the repeated sharing of misinformation items with the same source links and account for different types of user behavior. This chapter is based on the paper published in PACMHCI (CSCW 2025) [72].

[72]: Chuai, Zhao, and Lenzini (2025), 'From news source sharers to post viewers: How topic diversity and conspiracy theories shape engagement with misinformation during a health crisis'

Chapter 3 addresses **RO2** by examining how professional fact-checking organizations select political statements for verification. Focusing on statements that mention U.S. political figures, the chapter finds that fact-checked false statements are more likely than true ones to reference political elites. It also uncovers clear partisan asymmetries, with false statements disproportionately mentioning more Democrats than Republicans, and highlights differences in emotional framing between fact-checked true and false statements. These findings provide insights

Table 1.1: Thesis structure showing the correspondence between chapters, research objectives, and associated publications.

Chapter #	Research objectives	Associated publication
Part I		
Chapter 2	RO1	Y. Chuai, J. Zhao, and G. Lenzini, ‘From news source sharers to post viewers: How topic diversity and conspiracy theories shape engagement with misinformation during a health crisis,’ <i>Proceedings of the ACM on Human-Computer Interaction</i> , vol. 9, no. 7, pp. 1–53, 2025.
Chapter 3	RO2	Y. Chuai, J. Zhao, N. Pröllochs, and G. Lenzini, ‘Is fact-checking politically neutral? Asymmetries in how U.S. fact-checking organizations pick up false statements mentioning political elites,’ in <i>Proceedings of the International AAAI Conference on Web and Social Media</i> , vol. 19, 2025, pp. 403–429.
Part II		
Chapter 4	RO3	Y. Chuai, H. Tian, N. Pröllochs, and G. Lenzini, ‘Did the roll-out of Community Notes reduce engagement with misinformation on X/Twitter?’ <i>Proceedings of the ACM on Human-Computer Interaction</i> , vol. 8, no. CSCW2, pp. 1–52, 2024.
Chapter 5	RO3	Y. Chuai, M. H. Ribeiro, G. Lenzini, and N. Pröllochs, ‘When “likers” go private: Engagement with reputationally risky content on X,’ <i>ArXiv</i> , 2026.
Chapter 6	RO3	Y. Chuai et al., ‘Community-based fact-checking reduces the spread of misleading posts on X (formerly Twitter),’ <i>Nature Communications</i> , 2026, forthcoming.
Chapter 7	RO3	Y. Chuai, A. Sergeeva, G. Lenzini, and N. Pröllochs, ‘Community fact-checks trigger moral outrage in replies to misleading posts on social media,’ in <i>Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems</i> , 2025, pp. 1–23.
Part III		
Chapter 8	RO4	Y. Chuai, G. Lenzini, and N. Pröllochs, ‘Consensus stability of community notes on X,’ in <i>Proceedings of the ACM on Web Conference 2026</i> , 2026, pp. 8885–8896.
Chapter 9	RO4	Y. Chuai, S. Zhang, Z. Wang, X. Yi, M. Mosleh, and G. Lenzini, ‘Request a note: How the request function shapes X’s Community Notes system,’ in <i>Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems</i> , 2026, pp. 1–22.

into how the politicization of content can influence fact-checking priorities and practices. This chapter is based on the paper published in Proceedings of ICWSM 2025 [73].

Part II: Effectiveness of community-based fact-checking. Part II includes four chapters to address RO3, with multi-dimensional evaluations of community-based fact-checking system – X’s Community Notes.

Chapter 4 addresses RO3 by evaluating the initial effectiveness of X’s Community Notes program. Using large-scale data, this chapter examines whether introducing Community Notes program reduces overall engagement with misleading posts in terms of retweets and likes. While the feature increased the volume of fact-checks, particularly for high-profile users, it did not significantly lower overall engagement with misinformation. The findings suggest that, although promising, Community Notes system may act too slowly to influence content during its most viral stages, highlighting the importance of assessing fact-checking tools in real-world settings. This chapter is based on the paper published in PACMHCI (CSCW 2024) [74].

Chapter 5 provides a complementary analysis for RO3 by studying a confounding design change on X: hiding public visibility of likes on posts. This chapter investigates whether hiding who likes content affects user behavior, particularly for “high-exposure” posts that people might

[73]: Chuai, Zhao, Pröllochs, and Lenzini (2025), ‘Is fact-checking politically neutral? Asymmetries in how U.S. fact-checking organizations pick up false statements mentioning political elites’

[74]: Chuai, Tian, Pröllochs, and Lenzini (2024), ‘Did the roll-out of Community Notes reduce engagement with misinformation on X/Twitter?’

[75]: Chuai, Horta Ribeiro, Lenzini, and Pröllochs (2026), ‘When “likers” go private: Engagement with reputationally risky content on X’

[76]: Chuai, Pilarski, Renault, Restrepo-Amariles, Troussel-Clément, Lenzini, and Pröllochs (2026), ‘Community-based fact-checking reduces the spread of misleading posts on X (formerly Twitter)’

[77]: Chuai, Sergeeva, Lenzini, and Pröllochs (2025), ‘Community fact-checks trigger moral outrage in replies to misleading posts on social media’

[78]: Chuai, Lenzini, and Pröllochs (2026), ‘Consensus stability of community notes on X’

[79]: Chuai, Zhang, Wang, Yi, Mosleh, and Lenzini (2026), ‘Request a note: How the request function shapes X’s Community Notes system’

avoid liking due to social pressure. Combining a platform-level analysis and a user survey, the study finds limited changes in user engagement. While participants report a slight increase in willingness to like sensitive content, overall engagement remains largely unaffected. Therefore, these findings suggest that platform-level engagement signals are relatively robust to visibility changes, implying that the engagement metrics used to evaluate fact-checking effectiveness are driven more by content relevance and automated diffusion than by social signaling concerns. This chapter is based on a preprint available on arXiv [75].

Chapter 6 addresses **RO3** by examining how community-based fact-checking affects the *temporal dynamics* of engagement with misinformation. It shows that once a community note is displayed, the subsequent sharing of misleading posts drops substantially (on average, 61.2%), and authors are more likely to delete their posts. Importantly, the chapter finds that community notes often appear too late to intervene during the early and most viral stage of misinformation diffusion, which limits their overall platform-wide effectiveness. These findings highlight the importance of timing in fact-checking interventions and suggest that earlier or faster annotation mechanisms are crucial for reducing the spread of misinformation at scale. This chapter is based on the paper accepted at Nature Communications [76].

Chapter 7 addresses **RO3** by examining users’ emotional responses to community fact-checks. Using large-scale data from replies to fact-checked posts, the chapter finds that displaying community notes increases negative emotions such as anger, disgust, and moral outrage. These results suggest that users view misinformation as a violation of social norms and respond strongly when content is debunked. This chapter is based on the paper published in Proceedings of CHI 2025 [77].

Part III: Resilience and participation in community-based fact-checking. Part III includes two chapters to address **RO4** and assess the optimization and evolution of community-based fact-checking systems.

Chapter 8 addresses **RO4** by evaluating the resilience of Community Notes’ algorithm for selecting “helpful” notes. Using millions of ratings from diverse contributors, the chapter shows that about 30% of notes initially marked as helpful later lose this status due to post-display changes in ratings. It finds that contributors with differing viewpoints react differently: supporters increase positive ratings, while dissimilar users tend to give negative ratings, producing post-display polarization. These findings highlight the vulnerability of consensus-based systems to ideological bias and suggest strategies to improve the stability and reliability of crowdsourced fact-checking. This chapter is based on the paper accepted at The Web Conf 2026 [78].

Chapter 9 also addresses **RO4**, focusing on the “Request Community Note” feature, which allows users to request fact-checks for specific posts. This chapter shows that contributors prioritize posts that are estimated to be misleading and authored by users with high misinformation exposure, but often overlook political content highlighted by requestors. Notes generated through requests tend to be more helpful and less polarized than others, reflecting selective and high-quality participation from top contributors. These results highlight both the potential and the limitations of expanding user participation to scale crowdsourced fact-checking. This chapter is based on the paper accepted at CHI 2026 [79].

1.5 Summary of Main Contributions

This thesis makes several key contributions to the study of misinformation on social media and the evaluation of fact-checking interventions. These contributions span theoretical, methodological, and practical dimensions:

- ▶ **Advancing the understanding of user engagement with misinformation**
 - This thesis provides a nuanced analysis of how users interact with misinformation during health crises, focusing on multi-topic content and conspiracy narratives.
 - By distinguishing between news source sharers and post viewers, it reveals that engagement patterns differ across user levels: while sharers sustain misinformation over time, viewers are more influenced by topic diversity and conspiratorial framing.
 - These insights inform the design of targeted interventions that consider user heterogeneity, helping to mitigate the spread of misinformation more effectively.
- ▶ **Empirical insights into partisan asymmetries in fact-checking practices**
 - The thesis investigates how professional fact-checking platforms pick up claims mentioning political elites, uncovering partisan asymmetries and temporal patterns around election cycles.
 - These findings contribute to a better understanding of the interplay between misinformation, politicization, and fact-checking practices, highlighting potential areas for improving transparency and fairness in professional fact-checking.
- ▶ **Evaluating the effectiveness of community-based fact-checking**
 - Using large-scale, real-world data from X's Community Notes system, this research provides causal evidence on the impact of community-based fact-checks on user engagement, post deletion, and emotional reactions.
 - It demonstrates that while community notes can reduce the spread of misleading content and trigger moral accountability, their effectiveness is limited by delays in display and the influence of high-profile content.
 - These findings offer practical guidance for enhancing crowd-sourced fact-checking strategies and platform design to maximize timely intervention.
- ▶ **Assessments of resilience and evolution of community-based fact-checking**
 - The thesis examines the resilience of the Community Notes algorithm and the effects of expanded participation through the Request Community Note feature on X.
 - It identifies vulnerabilities, such as post-display polarization and selection biases, while also demonstrating that current request function is limited in expanding coverage.
 - These contributions provide actionable recommendations for designing more robust, scalable, and participatory fact-checking systems.

Overall, this thesis contributes to theory, method, and practice by: (i) deepening the understanding of misinformation diffusion and user behavior, (ii) uncovering the partisan asymmetries of fact-checking, (iii) evaluating the effectiveness of innovative, crowdsourced interventions, and (iv) offering system-level insights for improving community-based fact-checking. Collectively, these contributions advance both the scientific

study of misinformation and the practical design of interventions to safeguard information integrity on social media platforms.

1.6 List of Publications

The following is a list of all my first-author publications during the PhD studies. Eight of them are included in this thesis (see Table 1.1).

- [1] Y. Chuai, G. Lenzini, and N. Pröllochs, ‘Consensus stability of community notes on X,’ in *Proceedings of the ACM on Web Conference 2026*, 2026, pp. 8885–8896, [CORE A*; SnT Top 10].
- [2] Y. Chuai, M. H. Ribeiro, G. Lenzini, and N. Pröllochs, ‘When “likers” go private: Engagement with reputationally risky content on X,’ *ArXiv*, 2026.
- [3] Y. Chuai, S. Zhang, Z. Wang, X. Yi, M. Mosleh, and G. Lenzini, ‘Request a note: How the request function shapes X’s Community Notes system,’ in *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 2026, pp. 1–22, [CORE A*; SnT Top 10].
- [4] Y. Chuai et al., ‘Community-based fact-checking reduces the spread of misleading posts on X (formerly Twitter),’ *Nature Communications*, 2026, forthcoming.
- [5] Y. Chuai, A. Sergeeva, G. Lenzini, and N. Pröllochs, ‘Community fact-checks trigger moral outrage in replies to misleading posts on social media,’ in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025, pp. 1–23, [CORE A*; SnT Top 10].
- [6] Y. Chuai, J. Zhao, and G. Lenzini, ‘From news source sharers to post viewers: How topic diversity and conspiracy theories shape engagement with misinformation during a health crisis,’ *Proceedings of the ACM on Human-Computer Interaction*, vol. 9, no. 7, pp. 1–53, 2025, [CORE A; SnT Top 10].
- [7] Y. Chuai, J. Zhao, N. Pröllochs, and G. Lenzini, ‘Is fact-checking politically neutral? Asymmetries in how US fact-checking organizations pick up false statements mentioning political elites,’ in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 19, 2025, pp. 403–429, [CORE A; SnT Top 10].
- [8] Y. Chuai and G. Lenzini, ‘Changes in sentiments and user engagement for 2024 US presidential candidates after Biden’s withdrawal: An analysis of TikTok videos,’ *ArXiv*, 2024.
- [9] Y. Chuai, H. Tian, N. Pröllochs, and G. Lenzini, ‘Did the roll-out of Community Notes reduce engagement with misinformation on X/Twitter?’ *Proceedings of the ACM on Human-Computer Interaction*, vol. 8, no. CSCW2, pp. 1–52, 2024, [CORE A; SnT Top 10].
- [10] Y. Chuai, A. Rossi, and G. Lenzini, ‘Using emotions and topics to understand online misinformation,’ in *International Conference on Web Engineering*, 2023, pp. 395–400.

The following is a list of publications to which I contributed as a co-author during my PhD studies.

- [1] S. Zhang et al., 'Collab: Fostering critical identification of deepfake videos on social media via synergistic annotation,' in *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 2026, pp. 1–21, [CORE A*; SnT Top 10].
- [2] H. Tang, B. Rohles, Y. Chuai, G. Lenzini, and A. Sergeeva, 'Between privacy and transparency: Exploring the ways of communicating credibility assessment,' in *2025 European Symposium on Usable Security*, 2025, pp. 86–99.
- [3] Z. Wang et al., 'In a flap: Experiences with a bioinspired flying robot,' *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 9, no. 3, pp. 1–20, 2025.
- [4] S. Zhang et al., 'Commenotes: Synthesizing organic comments to support community-based fact-checking,' *ArXiv*, 2025.
- [5] D. Chapellier, Y. Chuai, J. De Meyere, L. Desaunettes, A. Sergeeva, and H. Tang, 'The potential impact of the DSA on disinformation,' in *Cyberspace Conference*, 2023.

PART I: MISINFORMATION DIFFUSION AND TRADITIONAL FACT-CHECKING

User Engagement With Misinformation During Health Crises

2

Abstract: Online engagement with misinformation threatens societal well-being, particularly during health crises when susceptibility to misinformation is heightened in a multi-topic context. Here, we focus on the COVID-19 pandemic and address a critical gap in understanding engagement with multi-topic misinformation on social media at two user levels: news source sharers (who post news items) and post viewers (who engage with news posts). To this end, we analyze 7273 fact-checked source news items and their associated posts on $\mathbb{X}$ through the lens of topic diversity and conspiracy theories. We find that false news, especially those containing conspiracy theories, exhibits higher topic diversity than true news. At news source sharer level, false news has a longer lifetime and receives more posts on $\mathbb{X}$ than true news, with conspiracy theories further extending its longevity. However, topic diversity does not significantly influence news source sharers' engagement. At post viewer level, contrary to news source sharer level, posts characterized by heightened topic diversity receive more reposts, likes, and replies. Notably, post viewers tend to engage more with misinformation containing conspiracy narratives: false news posts that contain conspiracy theories, on average, receive 40.8% more reposts, 45.2% more likes, and 44.1% more replies compared to those without conspiracy theories. Our findings suggest that news source sharers and post viewers exhibit distinct engagement patterns on $\mathbb{X}$, offering valuable insights into refining misinformation interventions at these two user levels.

Publication

Y. Chuai, J. Zhao, and G. Lenzini, 'From news source sharers to post viewers: How topic diversity and conspiracy theories shape engagement with misinformation during a health crisis,' *Proceedings of the ACM on Human-Computer Interaction*, vol. 9, no. 7, pp. 1–53, 2025.

2.1 Introduction

The spread of misinformation – *i.e.*, general false or misleading information created either intentionally or unintentionally – on social media platforms has emerged as one of the most significant global risks in recent years, posing far-reaching consequences for societies and individuals [1, 74]. Health misinformation, compared to misinformation in other domains, carries unique dangers due to its potential to cause harm to human bodies and lives [80]. This issue is particularly concerning during public health crises, such as the COVID-19 pandemic, when heightened emotions, uncertainty, and increased information consumption create fertile ground for the dissemination of false or misleading information [8, 10, 81–83]. Additionally, as a distinct type of misinformation, conspiracy theories are deliberately fabricated and attempt to explain the causes of significant public events as secret plots by powerful and malicious groups [39–42]. Throughout the COVID-19 pandemic, misinformation and conspiracy theories regarding the virus, its origin, and potential treatments or prevention strategies have spread widely on social media,

2.1	Introduction	15
2.2	Background and Related Work	18
2.2.1	Misinformation on Social Media	18
2.2.2	Misinformation During Public Events/Crises . . .	18
2.2.3	Engagement With COVID-19 Misinformation on Social Media .	19
2.2.4	Multi-Topic Context and Conspiracy Theories During COVID-19	20
2.3	Data and Methods	21
2.3.1	Dataset: MM-COVID . .	21
2.3.2	COVID-19 Topic Spectrum	21
2.3.3	Topic Diversity	23
2.3.4	COVID-19 Conspiracy Theories	24
2.3.5	User Engagement on $\mathbb{X}$.	25
2.3.6	Empirical Models	25
2.4	Empirical Results	28
2.4.1	Topic Diversity (RQ1.1 & RQ2.1)	28
2.4.2	News Source Sharer Engagement: News Lifetime and Post Count (RQ1.2 & RQ2.2)	30
2.4.3	Post Viewer Engagement: Reposts, Likes, and Replies (RQ1.2 & RQ2.2) .	31
2.4.4	Robustness Checks	33
2.4.5	Summary of Main Findings	34
2.5	Discussion	34
2.5.1	Relevance	34
2.5.2	Research Implications . .	35
2.5.3	Practical Implications . .	37
2.5.4	Limitations and Further Research	37
2.6	Conclusion	38

undermining pandemic control efforts [84–86]. For instance, widespread anti-vaccine conspiracy theories on social media have strengthened beliefs in the ineffectiveness and side effects of vaccinations, contributing to individuals' vaccination hesitancy and refusal, thereby increasing the risk of COVID-19 infection [2, 3, 85].

Notably, during health crises, various aspects of society become interconnected. The discussions traverse multiple domains and demonstrate strong interconnectedness, contributing to a rich and diverse informational landscape [37, 38]. For example, a comprehensive analysis of 13.9 million COVID-19 posts on ￼ (formerly Twitter) reveals discussions spanning more than 20 topics across society, encompassing areas such as politics, economy, and racism [87]. Yet, an understanding of how users engage with misinformation within the multi-topic context during health crises is still in its infancy. Vosoughi et al. [14] examined a large-scale of true and false news posts distributed on ￼ from 2006 to 2017, finding that false news posts are more novel and attract more replies expressing surprise compared to true news posts. The novelty in their study was measured by the topic similarity between the true (false) news post a user reposted and the previous posts to which the user had been exposed before the repost. However, other research shows that social media users often have specific topic preferences and tend to consume similar content, especially within groups of like-minded individuals, *i.e.*, echo chambers [29]. Additionally, social media recommendation algorithms frequently feed users content based on their view histories and increase echo chambers, further reinforcing users' preferences [88, 89]. Reconciling these seemingly conflicting findings, one possible explanation is that false news may strategically connect users' interests with other topics to pursue (hidden) agendas. Particularly, in the context of health crises marked by high uncertainty, diverse topics are more interconnected within discussions around crises [87]. Therefore, an analysis of topic interconnectedness in true and false news items during health crises is warranted.

Previous research has shown that the diversity of topics within information content (*i.e.*, multi-topic information) plays an important role in shaping perceptions and assessments of credibility through the mechanism of associative inference [90]. Associative inference is an adaptive process that enables individuals to link related information and form novel connections, even in the absence of direct experience, thereby increasing individuals' susceptibility to online misinformation [90]. Moreover, the associative inference around online misinformation can evolve within a reinforcing cycle. Individuals collaboratively construct a misleading version of reality and amplify alleged evidence. This facilitates subsequent actions that, in turn, further reinforce the manufactured reality [91]. Therefore, the interconnected and associative multi-topic discussions during health crises underscore the potential role of topic diversity in driving engagement with misinformation on social media. Based on this rationale, we address the following research questions to explore topic diversity in misinformation (RQ1.1) and examine whether it shapes user engagement (RQ1.2):

- ▶ RQ1.1: Does misinformation exhibit higher topic diversity compared to true information?
- ▶ RQ1.2: Is topic diversity positively associated with user engagement, and is this association stronger for misinformation than for true information?

Moreover, the integration of conspiracy theories within misinformation narratives serves as an effective strategy to deliberately interconnect

various topics [37, 42]. This deliberate approach contributes to an increase in individuals' beliefs in misinformation by fostering illusory pattern perception and promoting associative inference [90, 92]. In the early months of the COVID-19 pandemic, social media platforms were inundated with various conspiracy theories related to the origins, spread, and treatment of the virus [84]. The propagation of misinformation with coordinated conspiracies has promoted distrust in public health authorities and skepticism about the safety and effectiveness of COVID-19 vaccines [85, 93]. While the proliferation of conspiracy theories on social media has raised significant concerns, little is known about how the incorporation of conspiracy theories in misleading narratives differentiates individuals' engagement on social media compared to other types of misinformation [94]. Given this, we further investigate the following research questions to analyze whether misinformation containing conspiracy theories exhibits higher topic diversity (RQ2.1) and receives more engagement (RQ2.2) compared to misinformation without conspiracy theories:

- RQ2.1: Does misinformation containing conspiracy theories exhibit higher topic diversity compared to misinformation without conspiracy theories?
- RQ2.2: Does misinformation containing conspiracy theories receive more engagement compared to misinformation without conspiracy theories?

To address our research questions, we perform a comprehensive analysis of online engagement with multi-topic misinformation during the early stages of the COVID-19 pandemic, focusing on topic diversity and conspiracy theories. It is worth noting that we examine engagement at two levels: *news source sharer level* and *post viewer level*. News source sharers create posts on $\mathbb{X}$ that link to source news items, while post viewers encounter these news posts within the platform and then engage by reposting, liking, or replying. The two user groups may have different motivations for engagement [42, 86, 95]. Additionally, while previous studies have extensively explored post viewer engagement on social media [10, 14, 34, 81], research on news source sharer engagement remains limited. Understanding this under-investigated area is crucial to complement our knowledge of news-sharing behavior.

Methodology: We conduct our analysis based on a COVID-19 dataset consisting of 7273 fact-checked source news claims and their corresponding posts on $\mathbb{X}$. These source news claims span six languages and cover the initial period of COVID-19 from January to September 2020. At news source sharer level, the engagement metrics include the number of posts linking to a specific news item and the lifetime of the news item from the creation of its first post to its latest post. At post viewer level, the engagement metrics include repost count, like count, and reply count. To gain a comprehensive understanding of the topic spectrum surrounding COVID-19 discussions, we additionally use over 15 million daily sampled posts at the early stage of COVID-19 to train a Word2Vec model and collect topic keywords. Subsequently, we use the topic keywords to calculate topic diversity and identify conspiracy theories. Finally, we employ multiple regression models to investigate our research questions.

Contributions: Our work reveals that elevated topic diversity has been a characteristic of COVID-19 misinformation, particularly for those accompanied by conspiracy narratives. This characteristic becomes more significant within false news posts on $\mathbb{X}$ compared to their source news claims. At post viewer level, heightened topic diversity and the integration of conspiracy theories significantly correlate with increased engagement (*i.e.*, reposts, likes, and replies) with misinformation on $\mathbb{X}$. For instance,

false news posts that contain conspiracy theories, on average, receive 40.8% more reposts, 45.2% more likes, and 44.1% more replies compared to false news posts without conspiracy theories. However, at news source sharer level, topic diversity and conspiracy theories have no significant associations with the number of posts that source news items receive. This difference suggests that news source sharers and post viewers have distinct engagement patterns. Our findings offer valuable insights into understanding engagement with multi-topic misinformation from news source sharers to post viewers on social media during health crises. Additionally, by highlighting the different engagement patterns at news source sharer and post viewer levels in terms of topic diversity and conspiracy theories, this research informs the development of targeted interventions and strategies at both user levels to mitigate engagement with false information and foster a more resilient society.

2.2 Background and Related Work

2.2.1 Misinformation on Social Media

The spread of misinformation on social media strikes the stability of societies, and contributes to various contentious events in politics, finance, and pandemics [1]. Understanding how users engage with misinformation and contribute to its viral spread on social media can help social media platforms design targeted interventions. Prior research has shown that online misinformation spreads farther, faster, deeper, and more broadly compared to true information, as evidenced by the examination of online posts fact-checked by third-party organizations [14]. Nevertheless, another study, based on community fact-checked posts on $\mathbb{X}$, indicates that misinformation receives fewer reposts compared to true information [96]. This contradiction highlights that not every piece of misinformation spreads faster than true information. Instead, there are likely crucial factors behind the veracity of information that make misinformation unique and contribute to its online diffusion.

Several works have focused on understanding what characterizes misinformation and contributes to its viral spread. For instance, misinformation is more likely to incorporate negative emotions, such as anger, which has been identified as an important driver for information diffusion [10, 82, 97, 98]. Additionally, there is a partisan asymmetry that misinformation tends to be right-leaning and promote conservative positions [73]. Social media users are more likely to believe and share information that is congruent to their political ideologies [99–103]. Therefore, right-leaning users are more likely to share misinformation on social media, compared to left-leaning users [104]. More importantly, uncertain situations, such as public emergencies or crises, create fertile ground for online misinformation to thrive. During such times, people are eager to know more information about the situation and are thus more likely to believe and share related misinformation, which can undermine trust in control policies and cause detrimental harm to societies and individuals.

2.2.2 Misinformation During Public Events/Crises

During public events or crises, characterized by heightened emotions, uncertainty, and a surge in information consumption, falsehoods that cater to specific demands and offer related information to contextualize

the situation become rampant and easily spread [10]. The preservation of democracy is crucial during political elections. However, democracy can be undermined by the widespread dissemination of dedicated online misinformation aimed at pursuing specific political agendas. It has been found that people are exposed to misinformation more often during elections than usual [21, 73]. Election-related misinformation is often polarized, which can increase people's susceptibility and affect their voting behaviors [19, 104, 105]. Additionally, climate change poses challenges at multiple levels, with any mitigation efforts carrying significant economic and political implications [106]. Consequently, the debate surrounding global climate change has persisted for years and has become increasingly polarized on social media [35, 107, 108]. This polarized environment fosters climate change denial and misinformation, further complicating efforts to address and control climate change [35, 106].

The spread of misinformation during health crises is particularly severe, as it can directly harm human health and even cost lives [80]. Throughout the COVID-19 pandemic, misinformation on a wide variety of COVID-19 related topics has circulated widely on social media platforms. For example, the widespread misinformation on the COVID-19 vaccine reduces users' beliefs in its effectiveness, impeding the fight against the spread of COVID-19 [2, 3, 109]. A global survey has revealed that a substantial part of people consider COVID-19 misinformation as highly reliable [25]. Understanding how misinformation spreads and attracts users during such public events and crises is crucial for developing effective intervention strategies for emergency management.

2.2.3 Engagement With COVID-19 Misinformation on Social Media

The engagement with misinformation during health crises, particularly COVID-19, poses significant threats to crisis control and public safety. Several studies have scrutinized the factors that can influence engagement with COVID-19 misinformation on social media. Recognizing emotions as a significant driver of online information diffusion [10, 34, 110], research in this body has predominantly focused on the role of emotions during COVID-19 [8, 81, 82, 111]. For example, Solovev and Pröllochs [81] conducted a large-scale computational analysis on the role of moral emotions in the spread of COVID-19 misinformation. They found that false posts with higher other-condemning emotions (*e.g.*, anger) are more viral (*i.e.*, more reposts) than true posts. Based on experimental design, Freiling et al. [8] investigated the role of anxiety in misinformation belief and sharing during COVID-19, revealing that anxiety is a driving factor in believing and sharing both true and false claims. Wang et al. [112] put the focus on the images and found that online posts containing COVID-19 misinformation images receive similar interactions with posts containing random images. However, COVID-19 misinformation images are shared for longer periods than non-misinformation ones. Additionally, COVID-19 misinformation, mimicking the format and language features of news and scientific reports, is likely to receive more likes [113]. For the concern that social bots may share a large amount of misinformation during COVID-19, Teng et al. [86] analyzed different types of users in the engagement with misinformation. They found that the role of social bots in misinformation sharing is limited, and individuals with human-like behavior play a more prominent role.

Knowing the potential role of topics in engagement with misinformation, Ngai et al. [113] focused on the specific topics in the engagement with COVID-19 vaccine misinformation. They found that safety concerns are the most significant content theme but are negatively associated with likes and shares. However, during COVID-19, various aspects of society became interconnected. COVID-19 misinformation covered a broad spectrum of topics, including public health, politics, medicine, vaccines, and so forth [38, 87]. The multi-topic misinformation about COVID-19 carries high uncertainty and easily causes confusion to individuals, thus posing threats to people and society [38]. It has been revealed that individuals can adaptively link multiple related topics and make novel connections through the mechanism of associative inference, which increases the susceptibility to misinformation [90].

2.2.4 Multi-Topic Context and Conspiracy Theories During COVID-19

The discussion on COVID-19 during the pandemic not only focused on the direct health impacts but also spilled over into various social, economic, and political domains. Within the social domain, at the beginning of the pandemic, rumors and conspiracy theories about the origin of the virus in China were rampant on social media, exploiting anti-Asian racism and hate speech on social media [114]. Using COVID-19 as a weapon, malicious misinformation fostered online hate to spread quickly beyond the control of social media platforms [115]. The rising anti-Asian racism during the COVID-19 pandemic knocked Asian businesses [114, 116, 117]. In the economic domain, the pandemic hit the global economy, particularly impacting sectors such as travel and leisure [118]. Additionally, the scientific uncertainty surrounding the COVID-19 pandemic fueled the weaponization of COVID-19 for politicization [119]. Some individuals and groups have sought to exploit the pandemic to spread extremist ideologies and radicalize others [120]. The high degree of politicization within COVID-19 information contributed to polarization in attitudes towards the pandemic and even to political ideologies [121]. For instance, the debate over vaccination became increasingly polarized during the pandemic, expanding far beyond the medical discourse and entering the domain of organizational politics [122]. Taken together, the COVID-19 pandemic has been incorporated into many other societal domains to pursue specific interests, yielding significant spillover effects on the stability of societies. In this paper, we use topic diversity to denote the extent to which the information involves multiple topics and explore its impact on the engagement with misinformation on social media during COVID-19.

Particularly, at the onset of the COVID-19 pandemic, a myriad of conspiracy theories emerged, providing alternative explanations for the origins, spread, and treatment of the virus [84]. Examples include claims suggesting that the COVID-19 virus is a man-made bioweapon [84, 123]. Conspiracy theories often revolve around the notion of secret plots orchestrated by malevolent and powerful groups, deliberately crafted to achieve specific outcomes. This characteristic of conspiracy theories facilitates the connection of multiple topics through an illusory perception pattern, potentially fostering associative inference among individuals [37, 90, 92]. The integration of conspiracy theories within misinformation became more pronounced during the COVID-19 pandemic, marked by a significant increase in hyperlinks from misinformation websites

to conspiracy theories [42]. Despite this observation, the question of whether misinformation containing conspiracy theories receives more engagement than other types of misinformation remains to be explored. In light of these considerations, this paper tries to address the current research gaps by investigating the roles of topic diversity and conspiracy theories in the engagement with misinformation during health crises.

2.3 Data and Methods

2.3.1 Dataset: MM-COVID

The data for our analysis is sourced from a publicly available repository on GitHub [124]. This repository provides an extensive COVID-19 dataset that consists of 8825 source news claims and their corresponding engagement on $\mathbb{X}$. Initially, the repository gathers these news claims and their veracity verdicts from popular third-party fact-checking organizations worldwide, including Snopes and Poynter, as well as official health websites. To keep the quantity of the source claims in each language, the repository filters six languages: English, Spanish, Portuguese, Hindi, French, and Italian. This enables a broad and diverse examination of misinformation across linguistic and cultural contexts. Subsequently, posts that include these news sources are collected from $\mathbb{X}$ through its advanced search API. Notably, while the news sources in the dataset are linked to the original news URLs, precisely parsing the original news pages and obtaining their content are challenging due to the diverse range of news items collected from various languages and web pages. Instead, the dataset provides news claims, representing the main opinions and topics.

Data preprocessing: We begin by translating all the non-English texts in source news claims and their associated posts into English using Google Translate API via Googletrans.¹ Google Translate has a high performance in translating between English and non-English languages, retaining overall meaning in 82.5% of the translations [125, 126]. Subsequently, we perform text preprocessing by removing non-alphabetic and non-numerical characters, stop words, and URLs. We lemmatize all the words using the English transformer pipeline on Spacy and only consider news claims and posts that contain at least one lemmatized word. Additionally, we exclude posts that were posted before January 1, 2020. Due to the absence of release dates for source news items in the dataset, we use the first post that links to the specific source news item as its release date. In the end, we obtain 8822 news claims and associated 85,978 posts. Specifically, 59,363 posts are linked to 1543 false news claims, and 26,615 posts are linked to 7279 true news claims.

1: Googletans library is available on Github: <https://github.com/ssut/py-googletrans>. It is a free and unlimited Python library that implements Google Translate API and uses the same servers that translate.google.com uses.

2.3.2 COVID-19 Topic Spectrum

A previous study analyzed over 13 million English COVID-19 posts and identified 26 topics within 10 broad themes, including source, prevention, spread and growth, treatment and recovery, impact on the economy and markets, impact on health care sector, and government response, political impact, and racism [87]. This research also provided keywords corresponding to each topic. To ensure a comprehensive understanding of the topic spectrum surrounding COVID-19 discussions, we utilize these keywords as seeds to identify similar words and construct a topic

Table 2.1: COVID-19 topic spectrum. The 25 topics across 10 broad themes are included.

Theme	Topic
Source (origin)	Outbreak, alternative causes
Prevention	Social distancing, disinfecting and cleanliness
Symptoms	Symptoms
Spread and growth	Modes of transmission, spread of cases, death reports
Treatment and recovery	Drugs and vaccines, therapies, alternative methods, testing
Impact on health care sector	Impact on hospitals and clinics, health policy, frontline workers
Government response	Travel restrictions, financial measures, lockdown regulations
Impact on the economy and markets	Shortage of products, panic buying, stock markets, employment, impact on business
Political impacts	Political impacts
Racism	Racism

Table 2.2: The examples of topic seeds (bioweapon, vaccine, election, and quarantine) and their 10 most similar words.

bioweapon	vaccine	election	quarantine
biowarfare	vaccines	elections	isolation
biologicalweapon	vaccination	primaries	quarantined
bioweapons	vaccin	voting	selfquarantine
bioengineered	vax	2020election	quarentine
biologicalwarfare	vacine	electoral	selfisolation
bioterrorism	cure	election2020	quarantining
engineered	inoculation	referendum	quarantine
weaponized	treatments	vote	lockdown
wuhanlab	coronavirusvaccine	votes	quarintine
biowar	antibodies	generalelection	isolating

2: In a subsequent robustness check, we conduct the analysis again using all the original topics and keywords. The results are robust and consistently support our findings. The full estimation results are reported in Suppl. S2.3.

3: The dataset is available at <https://www.kaggle.com/datasets/smid80/coronavirus-covid19-tweets>.

4: The model is trained on the High Performance Computing platform at the University of Luxembourg.

lexicon based on a Word2Vec model. Upon reevaluating the associations between topic categories and their themes, we find that the keywords in the topic “Hotspots and location” relate solely to places, such as “cities” and “location,” not adequately reflecting the theme “Spread and growth” in terms of transmission modes or case numbers. Therefore, in the main analysis, we remove the topic “Hotspots and location” and focus on the remaining 25 topics across 10 themes (Table 2.1).²

Word2Vec model: To train the Word2Vec model, we additionally collect a large-scale COVID-19 dataset from a public repository on Kaggle.³ This dataset contains 26,854,078 posts in multiple languages, posted from March 9 to April 30, 2020 on X. Given the size of the dataset, we only filter English posts and do not involve the translation process. By combining this dataset with MM-COVID, we obtain a total of 15,059,238 English posts. After removing non-alphabetic and non-numerical characters as well as URLs in the posts, we utilize the refined corpus to train a Word2vec model with a vector size of 200 using the gensim library.⁴ The examples of the keywords and their ten most similar words are shown in Table 2.2.

Topic lexicon: We expand our set of topic keywords by incorporating words that exhibit cosine similarities of more than 0.6 with at least one of the initial topic seeds. This process yields 828 additional topic keywords, and they are automatically categorized into the same topics as their

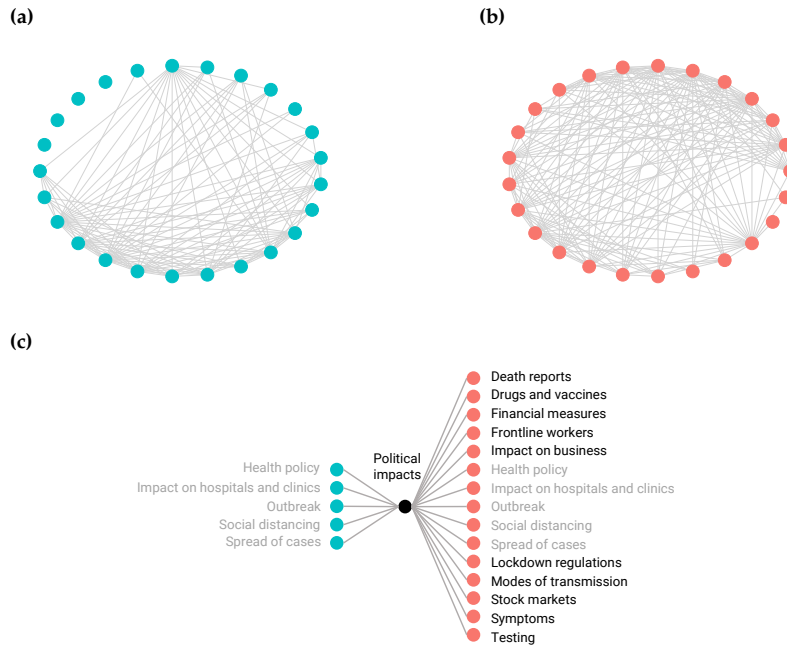


Figure 2.1: Topic co-occurrence in true and false news claims. (a) Topic co-occurrence network in true news claims. The number of edges is 107. (b) Topic co-occurrence network in false news claims. The number of edges is 167. (c) An example of topic co-occurrences for “Political impacts.” Shown are the topic “Political impacts” and its co-occurring topics in true and false news claims.

corresponding topic seeds. Additionally, we conduct a manual review of these keywords and remove ambiguous terms and abbreviations. For example, we remove the keyword “reason” from the topic “Alternative causes,” as it does not specify a particular alternative cause for COVID-19, whereas the keyword “chinvirus” is an effective term for this topic category. Similarly, we remove ambiguous abbreviations such as “pres” in the topic “Political impacts.” Ultimately, we retain 90.6% of the 828 keywords and integrate them with the topic seeds to create an extensive topic lexicon.⁵

Topic identification: We utilize the topic lexicon to identify topics discussed in both news claims and their associated posts. As a result, 82.4% of news claims contained at least one topic. Additionally, our analysis considers both the content of the source claims and the content of the corresponding posts when identifying topics in online posts. Consequently, 91.4% of posts contain at least one topic.

2.3.3 Topic Diversity

Topic diversity indicates the extent to which the content involves multiple topics. To give a sense of topic diversity, we construct topic co-occurrence networks for true and false news claims. Specifically, we calculate the frequency of topic co-occurrence of any two topics out of 25 topics in true and false news claims, and retain the co-occurrence edges with frequencies exceeding the median threshold. In the network of true news claims (Fig. 2.1a), the number of edges is 107, while in the network of false news claims (Fig. 2.1b), the number of edges is 167. Compared to true news claims, false news claims contain 56.1% more co-occurrence edges and are more interconnected, indicating higher interconnectedness and suggesting greater topic diversity in false news claims. For example, the topic “Political impacts” is connected to 5 topics in true news claims, whereas it connects to 10 additional topics, such as “Death reports” and “Drugs and vaccines,” in false news claims (Fig. 2.1c).

5: We repeat our analysis using the keywords without manual validation. The results remain consistent, further enhancing the robustness of our findings. The full estimation results are reported in Suppl. S2.3.

Table 2.3: COVID-19 conspiracy theories and corresponding keywords (phrases). “()” indicates that the word can be separated into a phrase. “{} and {}” indicates one of the words in the first set {} needs to appear with one of the words in the second set {}. “5G” and “Bill Gates” are case sensitive.

Conspiracy theory	Keywords
Origin	bat, batsoup, weapon, biowar, lab, conspiracy, chinese()virus, biochemical, bioterrorism, biowarfare, china()virus, wildlife, ccp()virus, chinese()flu, china()coronavirus, chinese()coronavirus, chinesevirus19, chines()virus, manmade, wuhan()coronavirus, wuhan()coronavirus, wuflu, wuhan()virus, chinaliedpeople
Vaccines	{vaccine, vax, vaccination, vaccin, vacine} and {anti, infertile, do not work, don't work, does not work, doesn't work, autism, autoimmune}
5G	5G
BigPharma	big pharma, fauci pharma, gates pharma
BillGates	Bill Gates
Exaggeration	does not exist, doesn't exist, exaggerated, inflated, refuse
FilmYourHospital	filmyourhospital, film your hospital, empty hospital, empty bed
GMO	gmo, genetically modified

Building upon previous research that used the Gini coefficient to measure the inequality of topic distribution in the content [37], we quantitatively measure topic diversity (*Diversity*) as:

$$Diversity = 1 - Gini = 1 - \frac{\sum_{i=1}^n \sum_{j=1}^n |x_i - x_j|}{2n^2 \bar{x}}, \quad (2.1)$$

where x_i denotes the number of words associated to topic i in each news claim (post), and n is the number of topics. In our analysis, we exclude news items and posts that do not relate to any topic, resulting in 7273 source news claims with corresponding 70,904 posts (see details in Table 2.4).

2.3.4 COVID-19 Conspiracy Theories

The outbreak of COVID-19 has fueled widespread conspiracy theories with misleading claims about its origin, treatment, and other related topics. We summarize the conspiracy theories identified in previous research [84, 85, 123] and use a set of keywords and phrases to identify conspiracy theories embedded in false news claims and subsequent posts. Specifically, we consider eight COVID-19 conspiracy theory categories that connect to at least one topic in our analysis: the sources and origins of the virus that caused the outbreak of COVID-19 (VirusOrigin), the connection between 5G and COVID-19 (5G), the role of big pharmaceutical companies (BigPharma) and Bill Gates (BillGates) in the spread of COVID-19, the claim that the scope of the pandemic was exaggerated (Exaggeration), the claim that hospitals were actually empty (FilmYourHospital), the idea that the genetically modified crops led to the emergence of COVID-19 (GMO), and the skepticism regarding the ineffectiveness and side-effects of vaccines (Vaccines).

We combine the conspiracy-specific terms adopted by prior studies [84, 123] and compile an extensive list of keywords and phrases to identify the eight conspiracy theories (Table 2.3). Subsequently, we match these keywords with the lemmatized words in each news claim or post, and for phrases, we directly match them with the original translated text.

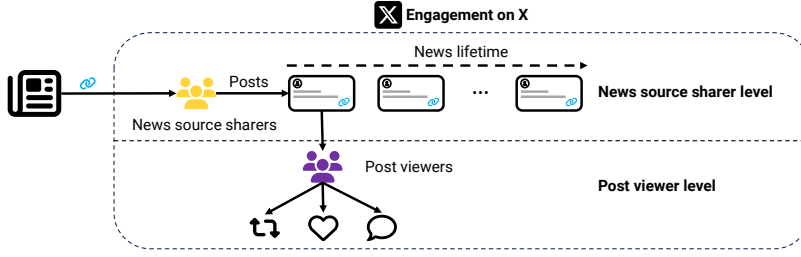


Figure 2.2: Engagement on X at news source sharer level and post viewer level. The news source sharers view original news items and share them on X through posts containing news links. Subsequently, post viewers engage with these news posts by reposting, liking, or replying.

We find that 89 false news claims and 4265 false news posts incorporate keywords (phrases) indicating conspiracy theories. Given that the shares of conspiracy posts range from 0.6% to 18% in previous studies [84], the 6% share of posts containing conspiracy theories in our dataset is reasonable. Upon a manual check of the 89 source news items that contain keywords or phrases related to conspiracy theories, we find that the accuracy is 92.1%. Therefore, the selected keywords (phrases) can effectively identify conspiracy theories in our dataset.

2.3.5 User Engagement on X

On the platform of X, users can interact with posts they view through reposting, liking, or replying. Existing research typically uses the number of reposts to study engagement with misinformation on social media [10, 14, 34, 74, 96]. Additionally, the number of likes and the number of replies are also important indicators of engagement with online posts [15, 74]. However, few studies examine engagement with misinformation from the perspective of news source sharing. Multiple posts on X can link to the same source news items, which increases the visibility of the news items. In this paper, we consider user engagement at two levels: news source sharer level and post viewer level (Fig. 2.2). At news source sharer level, news items are shared by news source sharers via posts on X. The engagement metrics include the number of posts and the lifetime of news items from the creation of the first post to the creation of the latest post. At post viewer level, engagement is measured by how viewers interact with these posts. Post viewers can repost, like, and reply to posts that are created by news source sharers and contain links to the original news items.

2.3.6 Empirical Models

We employ multiple explanatory regression models to explore our research questions separately. To this end, we define the following variables to measure *engagement at news source sharer level and post viewer level*.

- The engagement at news source sharer level:
 - *NewsLifetime*: The duration (in hours) from the creation of the first post to the creation of the latest post on X for the specific news item.
 - *NewsPostCount*: A count variable indicating the number of posts associated with the specific news item.
- The engagement at post viewer level:
 - *RepostCount*: A count variable indicating the number of reposts received by the specific post.

- *LikeCount*: A count variable indicating the number of likes received by the specific post.
- *ReplyCount*: A count variable indicating the number of replies received by the specific post.

Additionally, we define the following *key variables* to address our research questions:

- ▶ *Falsehood*: A dummy variable indicating whether the news claim is false (= 1) or true (= 0).
- ▶ *Conspiracy*: A dummy variable indicating whether the false news claim or post contains conspiracy theories (= 1) or not (= 0).
- ▶ *Diversity*: The topic diversity of the news claim or post.

We also control for the variables from *content characteristics* that can affect user engagement:

- ▶ *Words*: A continuous-like count variable indicating the number of words in the text of the news claim or post. This control variable is commonly used to account for the content richness [37].
- ▶ *Media*: A dummy variable indicating whether the post has media elements (= 1) or not (= 0). Previous research has found that media elements, such as images and videos, can facilitate user engagement [127].
- ▶ Emotions are considered to be the main driver of online information diffusion [10, 34, 110]. Our analysis considers six cross-cultural emotions – anger, disgust, fear, joy, sadness, and surprise – out of the eight basic emotions [10, 128]. We employ the commonly used NRC emotion lexicon to count emotion words in news claims and posts. Notably, 92.1% news claims and 95% posts contain at least one emotion word. Subsequently, we calculate the weight of each emotion in the specific text content based on the number of corresponding emotion words. If the news claims (posts) have no emotion words, all the weights of emotions are set to 0.

Finally, the posters on social media play an important role in driving engagement with their posts [10, 14]. We use the following four variables to account for *poster characteristics*:

- ▶ *Verified*: A dummy variable indicating whether the post account is verified (= 1) or not (= 0).
- ▶ *AccountAge*: A continuous-like count variable indicating the age (in days) of the post account.
- ▶ *Followers*: A continuous-like count variable indicating the number of followers of the post account.
- ▶ *Followees*: A continuous-like count variable indicating the number of followees of the post account.

The descriptive statistics for the above variables are shown in Table 2.4.

Topic diversity (RQ1.1 & RQ2.1): We use *Diversity* as the dependent variable to analyze topic diversity in true and false news from original news claims to online posts. In terms of claims, the linear regression model that incorporates the main independent variables of *Falsehood* and *Conspiracy* is specified as:

$$\begin{aligned} Diversity_i = & \beta_0 + \beta_1 Falsehood_i + \beta_2 Conspiracy_i + \beta_3 Words_i \\ & + \alpha' Emotions_i + u_{lang} + u_{time}, \end{aligned} \quad (2.2)$$

where *Emotions* indicates six basic emotions. u_{lang} denotes language-specific fixed effects, and u_{time} denotes month-year fixed effects.

	(1) All	(2) CTs
News Count	7,273	82
Falsehood	18.9%	100%
<u>Engagement metrics</u>		
Lifetime	315.617 (900.617)	1,968.141 (1,626.337)
Post Count	70,904	4,265
Repost Count	1,300,597	161,823
Like Count	3,943,848	452,106
Reply Count	380,224	107,363
<u>News characteristics</u>		
Dates	01/02/20 – 09/14/20	01/07/20 – 09/08/20
Diversity	0.083 (0.045)	0.089 (0.038)
Words	15.614 (6.364)	11.841 (6.739)
Anger	0.058 (0.109)	0.040 (0.093)
Disgust	0.045 (0.108)	0.059 (0.153)
Fear	0.183 (0.208)	0.149 (0.183)
Joy	0.065 (0.128)	0.097 (0.215)
Sadness	0.096 (0.139)	0.089 (0.135)
Surprise	0.045 (0.099)	0.026 (0.080)
<u>Post characteristics</u>		
Dates	01/02/20 – 09/14/20	01/09/20 – 09/09/20
Diversity	0.104 (0.052)	0.111 (0.050)
Words	26.372 (10.192)	27.006 (8.876)
Anger	0.068 (0.100)	0.069 (0.090)
Disgust	0.074 (0.134)	0.064 (0.105)
Fear	0.195 (0.177)	0.157 (0.146)
Joy	0.062 (0.106)	0.088 (0.163)
Sadness	0.129 (0.133)	0.095 (0.106)
Surprise	0.052 (0.091)	0.050 (0.088)
Media	21.9%	13.6%
Verified	18.5%	7.4%
AccountAge	2,460.659 (1,457.714)	2,259.347 (1,393.434)
Followers	369,041.439 (1,755,545.224)	56,847.547 (1,399,857.795)
Followees	3,031.517 (15,099.135)	2,361.965 (15,535.361)

Table 2.4: Dataset overview. Reported are mean values or count numbers for the variables (standard deviations in parentheses). Column (1) includes all the news claims and associated posts. 51.7% in English, 21.6% in Spanish, 4.9% in Portuguese, 15.7% in Hindi, 3.2% in French and 3.0% in Italian. Column (2) includes false news claims (posts) that contain conspiracy theories.

Given that each news item can have multiple posts, we need to control for the possible heterogeneity among news items when analyzing the topic diversity in terms of online posts. Therefore, we specify a liner regression model incorporating news-specific random effects:

$$\begin{aligned}
 Diversity_i = & \beta_0 + \beta_1 Falsehood_i + \beta_2 Conspiracy_i + \beta_3 Words_i \\
 & + \beta_4 Media_i + \beta_5 Verified_i + \beta_6 AccountAge_i \\
 & + \beta_7 Followers_i + \beta_8 Followees_i + \alpha' Emotions_i \\
 & + u_{lang} + u_{time} + v_{news},
 \end{aligned} \tag{2.3}$$

where v_{news} indicates news-specific random effects. Additionally, we incorporate *Media* and poster characteristics as control variables. In the above two models, the coefficient estimate of *Falsehood* (β_1) is to address RQ1.1, and the coefficient estimate of *Conspiracy* (β_2) is to address RQ2.1.

Engagement at news source sharer level – news lifetime and post count (RQ1.2 & RQ2.2): The first layer of engagement is from source news claims to news posts on $\mathbb{X}$. We examine how topic diversity and conspiracy theories influence engagement with misinformation in terms of news lifetime and post count. First, we take *NewsLifetime* as the dependent variable and specify a linear regression model as follows:

$$\begin{aligned}
 Lifetime_i = & \beta_0 + \beta_1 Falsehood_i + \beta_2 Conspiracy_i + \beta_3 Diversity_i \\
 & + \beta_4 Falsehood_i \times Diversity_i + \beta_5 Words_i + \alpha' Emotions_i \\
 & + u_{lang} + u_{time},
 \end{aligned} \tag{2.4}$$

where the interaction term of *Falsehood* $\times$ *Diversity* is to estimate the moderating effect of *Falsehood* on *Diversity*. Subsequently, we use a negative binomial regression model to explain *PostCount*. The dependent variable is $\log(E(PostCount_i|x_i))$, and the independent variables remain consistent with Eq. (2.4).

Engagement at post viewer level – reposts, likes, and replies (RQ1.2 & RQ2.2): To examine the effects of topic diversity and conspiracy theories on social engagement (*i.e.*, reposts, likes, and replies) with misinformation on $\mathbb{X}$, we specify a negative binomial regression model with language-specific fixed effects, month-year fixed effects, and news-specific random effects:

$$\begin{aligned} \log(E(y_i|x_i)) = & \beta_0 + \beta_1 Falsehood_i + \beta_2 Conspiracy_i + \beta_3 Diversity_i \\ & + \beta_4 Falsehood_i \times Diversity_i + \beta_5 Words_i + \beta_6 Media_i \\ & + \beta_7 Verified_i + \beta_8 AccountAge_i + \beta_9 Followers_i \\ & + \beta_{10} Followees_i + \alpha' Emotions_i \\ & + u_{lang} + u_{time} + v_{news}, \end{aligned} \quad (2.5)$$

where x_i indicates all the independent variables, and y_i signifies the dependent variables, including *RepostCount_i*, *LikeCount_i* and *ReplyCount_i*. In the above models for engagement, the coefficient estimates of *Falsehood* (β_1), *Diversity* (β_3), and *Falsehood* $\times$ *Diversity* (β_4) are to address RQ1.2, and the coefficient estimate of *Conspiracy* (β_2) is to address RQ2.2. All the continuous variables in these models are z-standardized.

2.4 Empirical Results

2.4.1 Topic Diversity (RQ1.1 & RQ2.1)

Descriptive statistics. We start by examining topic diversity from news claims to news posts (Figs. 2.3a–2.3c). Fig. 2.3a shows that false news claims (mean of 0.086) exhibit 4.9% higher topic diversity compared to true news claims (mean of 0.082, $KS = 0.072$, $p < 0.001$). This difference is significantly magnified (43.0% higher) at the post level, where the average topic diversity in false news posts is 0.113, compared to 0.079 in true news posts (Fig. 2.3b, $KS = 0.333$, $p < 0.001$). Additionally, we examine the paired changes in topic diversity from claims to their corresponding posts (Fig. 2.3c), finding that topic diversity in false news items (increase of 0.031) increases 158.3% more than that in true news items (increase of 0.012, $KS = 0.396$, $p < 0.001$). For false news items containing conspiracy theories (CTs), the topic diversity at the claim level has no statistically significant difference compared to the overall false news items (Fig. 2.3a, $KS = 0.138$, $p = 0.095$). However, Fig. 2.3c indicates that the most significant increases in topic diversity from claims to posts are more likely to be in CT news items, compared to overall false news items ($KS = 0.044$, $p < 0.001$). This may lead to the significant difference in topic diversity between CT news and general false news at the post level (Fig. 2.3b, $KS = 0.036$, $p < 0.001$).

Regression results. We further conduct regression analysis to estimate the effects of *Falsehood* and *Conspiracy* on topic diversity (Figs. 2.3d–2.3f). In Fig. 2.3d, the coefficient estimate of *Falsehood* is significantly positive (*coef.* = 0.388, $p < 0.001$). This implies that the topic diversity in false news is, on average, 0.388 standard deviations higher than in true news at the claim level, consistent with the observations in Fig. 2.3a.

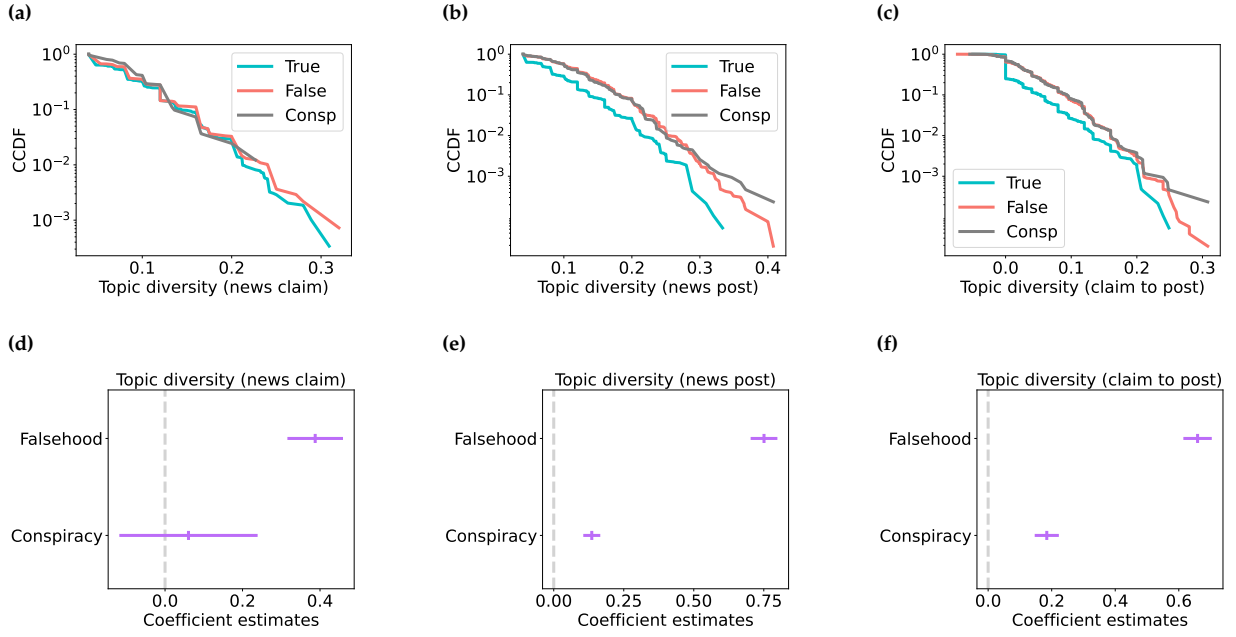
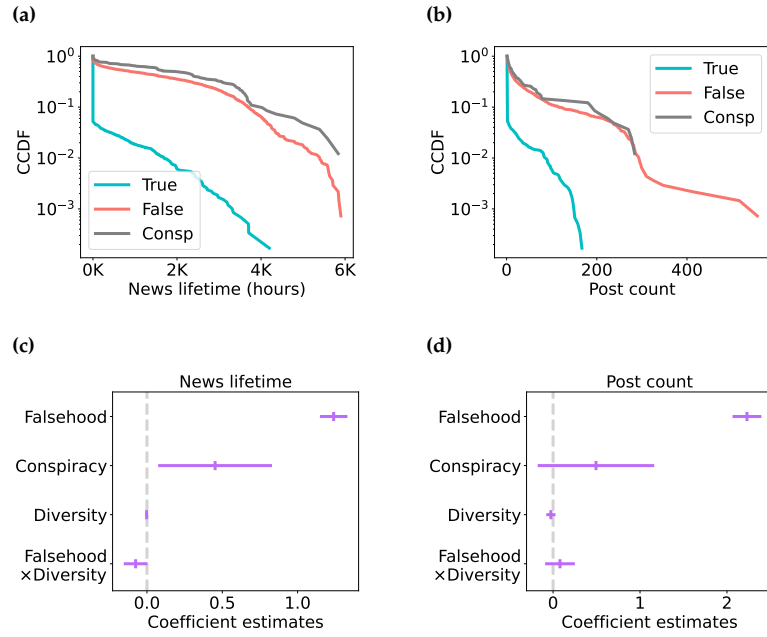


Figure 2.3: The topic diversity from news claims to news posts. (a)–(c) The Complementary Cumulative Distribution Functions (CCDFs) for (a) topic diversity at the source claim level, (b) topic diversity at the post level, and (c) the changes in topic diversity from claims to posts. Three categories of CCDFs are included in each figure, *i.e.*, true news (True), false news (False), and CT news (Consp). (d)–(f) The coefficient estimates of the independent variables – *Falsehood* and *Conspiracy*. Other independent variables are included during estimation but omitted in the visualization for better readability. Shown are mean values with error bars representing 95% Confidence Intervals (CIs). The dependent variables are (d) topic diversity in news claims, (e) topic diversity in news posts, and (f) the changes in topic diversity from claims to posts, respectively. See Table S2.1 in Suppl. S2.1 for full estimation results.

Additionally, the coefficient estimate of *Conspiracy* is not statistically significant, which indicates that the topic diversity in false news claims containing conspiracy theories is not significantly different from that in false news claims without conspiracy theories (Fig. 2.3d). Subsequently, we explore the topic diversity between true and false news at the post level. In Fig. 2.3e, the coefficient estimate of *Falsehood* is significantly positive ($coef. = 0.751, p < 0.001$), suggesting that the topic diversity in false news is, on average, 0.751 standard deviations higher than in true news at the post level.

Notably, as observed in the CCDF in Fig. 2.3b and the coefficient estimates of *Falsehood* in Figs. 2.3d–2.3e, the difference between true and false news posts is larger than that between true and false news claims. Additionally, the coefficient estimate of *Conspiracy* in Fig. 2.3e becomes significantly positive ($coef. = 0.136, p < 0.001$). This suggests that false news posts containing conspiracy theories exhibit even higher topic diversity (0.136 standard deviations more) than false news posts without conspiracy theories. There are two possible reasons for the larger difference in topic diversity between true and false news at the post level compared to the claim level. One is that news claims with higher topic diversity receive more posts on $\mathbb{X}$. Another is that false news has a bigger increase in topic diversity from the claim level to the post level, compared to true news. Given the second reason, we further examine the paired changes in topic diversity from the source news claims to the corresponding posts. In Fig. 2.3f, the positive coefficient estimate of *Falsehood* is statistically significant ($coef. = 0.659, p < 0.001$). This suggests that the increase in topic diversity in false news from the claim level to the post level is significantly larger (0.659 standard deviations more) than in true news. This finding could partially explain the enlarged difference in topic

Figure 2.4: The engagement at news source sharer level – news lifetime and post count. (a)–(b) The CCDFs for (a) news lifetime and (b) post count. Three categories of CCDFs are included in each figure, *i.e.*, true news (True), false news (False), and CT news (Consp). (c)–(d) The coefficient estimates of the independent variables – *Falsehood*, *Conspiracy*, *Diversity*, and *Falsehood*×*Diversity*. Other independent variables are included during estimation but omitted in the visualization for better readability. Shown are mean values with error bars representing 95% CIs. The dependent variables are (c) news lifetime and (d) post count, respectively. See Table S2.2 in Suppl. S2.1 for full estimation results.



diversity between true and false news posts, compared to their source claims. In addition, the coefficient estimate of *Conspiracy* in Fig. 2.3f is significantly positive (*coef.* = 0.185, $p < 0.001$). This means that the topic diversity in false news that contains conspiracy theories increases 0.185 standard deviations more than that in false news without conspiracy theories from the claim level to the post level. This finding could partially explain why the coefficient estimate of *Conspiracy* in Fig. 2.3d is not statistically significant, whereas it becomes significant in Fig. 2.3e.

In summary, we find that false news claims exhibit higher topic diversity compared to true news claims, and this difference is magnified at the post level (RQ1.1). Additionally, false news posts containing conspiracy theories have higher topic diversity than those without conspiracy theories, while no significant difference is observed at the claim level (RQ2.1).

2.4.2 News Source Sharer Engagement: News Lifetime and Post Count (RQ1.2 & RQ2.2)

Descriptive statistics. In terms of engagement at news source sharer level, we examine news lifetime that indicates the duration (in hours) from the creation time of the first post to the creation time of the last post for each news item on X. Fig. 2.4a shows that false news items have significantly longer lifetimes (mean of 1463.757 hours) than true news items (mean of 47.951 hours, $KS = 0.805$, $p < 0.000$). CT news items have longer lifetimes (mean of 1968.141 hours) than false news items ($KS = 0.167$, $p < 0.05$). Additionally, Fig. 2.4b shows that false news claims receive more posts (mean of 38.015) than true news claims (mean of 3.159, $KS = 0.806$, $p < 0.001$). However, the number of posts received by CT news items (mean of 44.976) does not differ significantly from that received by false news items ($KS = 0.113$, $p = 0.257$).

Regression results. We conduct regression analysis to further rigorously examine the effects of *Falsehood*, *Conspiracy*, and *Diversity* on news lifetime and post count (Figs. 2.4c–2.4d). In Fig. 2.4c, the coefficient estimate of *Falsehood* is significantly positive (*coef.* = 1.239, $p < 0.001$).

This suggests that, on average, false news items have lifetimes of 1.239 standard deviations longer for receiving posts on $\mathbb{X}$ compared to true news items. Additionally, the coefficient estimate of *Conspiracy* is also significantly positive (*coef.* = 0.452, $p < 0.05$). This means that, on average, the lifetimes for false news items containing conspiracy theories are 0.452 standard deviations longer than that for false news items without conspiracy theories. However, the coefficient estimates for *Diversity* and *Falsehood* $\times$ *Diversity* are not statistically significant. This suggests that topic diversity has no significant impact on the lifetimes of news items, irrespective of their veracity.

Additionally, in Fig. 2.4d, the coefficient estimate of *Falsehood* is significantly positive (*coef.* = 2.230, $p < 0.001$), indicating that the number of posts referencing a specific source news item is significantly associated with its veracity. On average, false news items are linked by $e^{2.230} - 1 = 8.3$ times more posts on $\mathbb{X}$ compared to true news items. Notably, the coefficient estimate of *Conspiracy* is not statistically significant for post count, which implies that the integration of conspiracy theories has no additional impact on the number of posts received by false news items. The coefficient estimate of *Diversity* is not statistically significant. It suggests that the topic diversity of a source news claim has no significant association with the number of online posts it receives. This finding does not support the first reason for the enlarged difference in topic diversity between true and false news posts, compared to that between true and false news claims (see Section 2.4.1). Furthermore, the coefficient estimate of the interaction term *Falsehood* $\times$ *Diversity* is also not statistically significant, reinforcing the finding that topic diversity has no significant effect on post count, irrespective of the veracity of news claims.

In summary, we find that, at news source sharer level, misinformation with higher topic diversity does not receive more engagement in terms of news lifetime and post count (RQ1.2). However, false news items containing conspiracy theories tend to have significantly longer lifetimes compared to false news items without conspiracy theories but do not receive a higher number of posts on $\mathbb{X}$ (RQ2.2).

2.4.3 Post Viewer Engagement: Reposts, Likes, and Replies (RQ1.2 & RQ2.2)

Descriptive statistics. In terms of engagement at post viewer level, we examine the numbers of reposts, likes, and replies for news posts. To this end, we first conduct descriptive analysis without incorporating the defined control variables from content and poster characteristics. We find that, on average, true news posts receive more engagement on $\mathbb{X}$ than false news posts in terms of reposts (Fig. 2.5a; true: 29.396, false: 14.403, $KS = 0.240$, $p < 0.001$), likes (Fig. 2.5b; true: 84.765, false: 45.234, $KS = 0.235$, $p < 0.001$), and replies (Fig. 2.5c; true: 6.294, false: 5.031, $KS = 0.133$, $p < 0.001$). However, it is worth noting that news posts that receive the most significant reposts, likes, or replies belong to the false category. Additionally, CT news posts receive more reposts (mean of 37.942, $KS = 0.043$, $p < 0.001$) and replies (mean of 25.173, $KS = 0.031$, $p < 0.01$) than overall false news posts, while without statistically significant difference in like count ($KS = 0.017$, $p = 0.194$).

Regression results. Incorporating multiple control variables related to content and poster characteristics that can affect engagement with news posts, we conduct regression analysis to estimate the effects of *Falsehood*,

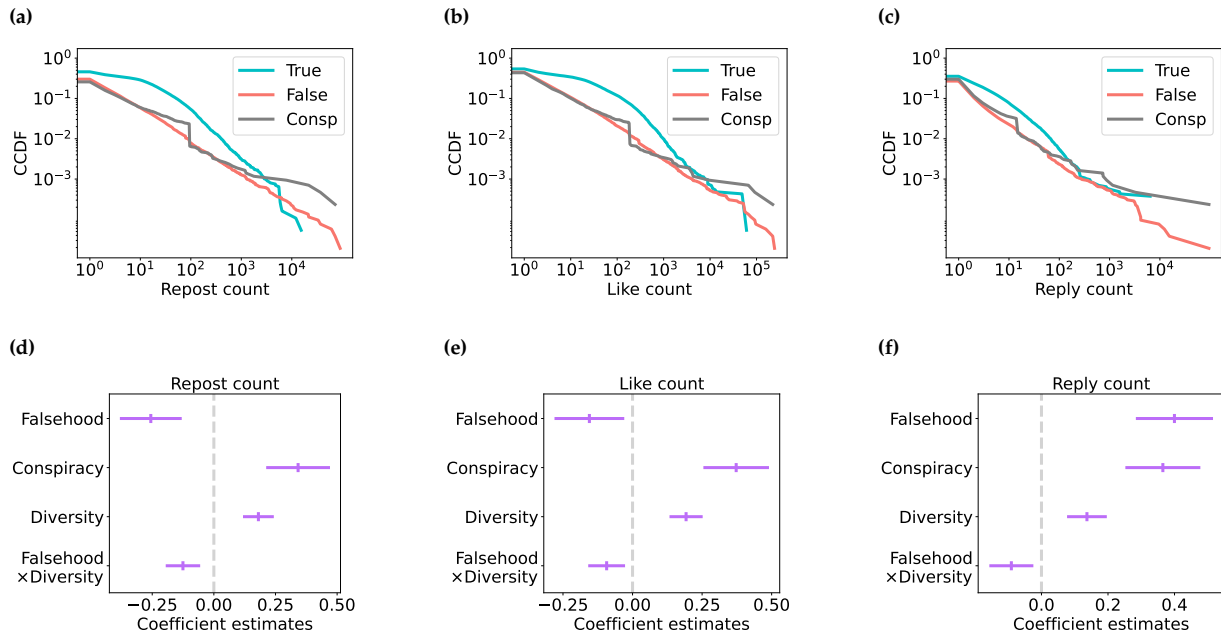


Figure 2.5: The engagement at post viewer level – repost count, like count, and reply count. (a)–(c) The CCDFs for (a) repost count, (b) like count, and (c) reply count. Three categories of CCDFs are included in each figure, *i.e.*, true news (True), false news (False), and CT news (Consp). (d)–(f) The coefficient estimates of the independent variables – *Falsehood*, *Conspiracy*, *Diversity*, and *Falsehood × Diversity*. Other independent variables are included during estimation but omitted in the visualization for better readability. Shown are mean values with error bars representing 95% CIs. The dependent variables are (d) repost count, (e) like count, and (f) reply count, respectively. See Table S2.3 in Suppl. S2.1 for full estimation results.

Conspiracy, and *Diversity* on the numbers of reposts, likes, and replies (Figs. 2.5d–2.5f). In Fig. 2.5d, the coefficient estimate of *Falsehood* is significantly negative ($coef. = -0.256, p < 0.001$), which means that false news posts receive fewer reposts than true news posts. On average, false news posts receive 22.6% fewer reposts than true news posts. Similarly, the coefficient estimate of *Falsehood* in Fig. 2.5e is significantly negative ($coef. = -0.155, p < 0.05$). This means that, on average, false news posts receive 14.4% fewer likes than true news posts. However, the coefficient estimate of *Falsehood* in Fig. 2.5f is significantly positive ($coef. = 0.400, p < 0.001$). This means that, on average, false news posts receive 49.2% more replies than true news posts.

Additionally, the coefficient estimates of *Diversity* in Fig. 2.5d ($coef. = 0.181, p < 0.001$), Fig. 2.5e ($coef. = 0.192, p < 0.001$), and Fig. 2.5f ($coef. = 0.137, p < 0.001$) are significantly positive. This indicates that heightened topic diversity is associated with increased online engagement in reposts, likes, and replies. Specifically, a one standard deviation increase in topic diversity is associated with 19.8% more reposts, 21.2% more likes, and 14.7% more replies. However, the coefficient estimates of *Falsehood × Diversity* in Fig. 2.5d ($coef. = -0.125, p < 0.001$), Fig. 2.5e ($coef. = -0.093, p < 0.01$), and Fig. 2.5f ($coef. = -0.091, p < 0.01$) are significantly negative.⁶ This suggests that the positive association of topic diversity with engagement on $\mathbb{X}$ is not significantly stronger within the context of false news posts compared to true news posts. On the contrary, the coefficient estimates of *Conspiracy* in Fig. 2.5d ($coef. = 0.342, p < 0.001$), Fig. 2.5e ($coef. = 0.373, p < 0.001$), and Fig. 2.5f ($coef. = 0.365, p < 0.001$) are significantly positive. This means that the integration of conspiracy theories within false news posts is positively linked to engagement at post viewer level: false news posts that contain conspiracy theories receive 40.8% more reposts, 45.2% more likes, and 44.1% more

6: The coefficient estimates of *Diversity* across repost count, like count, and reply count remain significantly positive in the subset analysis for false news posts. See Table S2.4 in Suppl. S2.1 for full estimation results.

replies compared to false news posts without conspiracy theories.

Clarification for the discrepancies between descriptive statistics and regression results: We observe two discrepancies between descriptive analysis and regression analysis: (i) In the descriptive analysis, Fig. 2.5c shows that true news posts receive more replies than false news posts, whereas Fig. 2.5f indicates that false news posts receive more replies than true news posts. (ii) No significant difference between CT news posts and overall false news posts is observed in the CCDF of like count (Fig. 2.5b), while Fig. 2.5e indicates that false news posts containing conspiracy theories receive more likes than those without conspiracy theories. These discrepancies suggest that the falsehood may correlate with other variations in the control variables, potentially complicating the findings in the descriptive analysis and underscoring the necessity of incorporating control variables that can affect engagement.

For instance, emotions, included as control variables in our regression models, have been well documented as a significant driver in facilitating online post sharing [10, 14, 81, 82, 97, 98]. We find that falsehood is positively correlated with anger ($r = 0.038, p < 0.001$) and surprise ($r = 0.051, p < 0.001$) but negatively correlated with joy ($r = -0.020, p < 0.001$; see Table S2.6 in Suppl. S2.2). This suggests that false news posts express more anger and surprise but less joy compared to true news posts. Within the estimation results of our regression models, the coefficient estimates of *Anger* for repost count ($coef. = 0.049, p < 0.001$), like count ($coef. = 0.059, p < 0.001$), and reply count ($coef. = 0.040, p < 0.01$) are consistently positive and statistically significant (see Table S2.3 in Suppl. S2.1 for full estimation results). Similarly, the coefficient estimates of *Surprise* for repost count ($coef. = 0.093, p < 0.001$), like count ($coef. = 0.120, p < 0.001$), and reply count ($coef. = 0.121, p < 0.001$) are also consistently positive and statistically significant. Our results suggest that news posts expressing higher levels of anger or surprise tend to receive more reposts, likes, and replies, aligning with findings from previous research [10, 14]. However, the coefficient estimate of *Joy* is not statistically significant for repost count, while significantly positive for reply count ($coef. = 0.077, p < 0.001$). This suggests that news posts expressing more joy are more likely to receive replies but do not receive more reposts. In light of this, we find two opposite effects: falsehood contributes to an increase in replies for false news posts, while the lower expression of joy in false news posts, compared to true news posts, links to fewer replies. The aggregation of these two opposite effects in the descriptive analysis could partially explain the discrepancy observed between Fig. 2.5c and Fig. 2.5f.

In summary, at post viewer level, news posts with higher topic diversity receive more reposts, likes, and replies; however, this association is not stronger for false news posts (RQ1.2). Additionally, false news posts containing conspiracy theories receive more reposts, likes, and replies compared to false news posts without conspiracy theories (RQ2.2).

2.4.4 Robustness Checks

To ensure the robustness of our findings, we conduct several additional checks: (i) We examine the variance inflation factors (VIFs) among the independent variables at both news claim and news post levels (see Table S2.7 in Suppl. S2.2). All the VIFs are significantly lower than the critical threshold of 4, which indicates that multicollinearity is not an issue for our analysis. (ii) To test whether our findings are robust with potential noise in topic keywords, we conduct our analysis again

7: QAnon is a far-right conspiracy theory alleging that a cabal of politicians and media personalities are engaged in a war to undermine society [94, 129–131]. The QAnon discourse has become more diverse and evolved into a global phenomenon during the COVID-19 pandemic [132]. We adopt a list of QAnon-specific keywords from previous research [133] to identify QAnon conspiracy theory in our dataset. The performance of this keyword list in identifying QAnon discourse has been manually validated, showing high agreement with human assessments [130]. As a result, we identify 6 QAnon news claims and 302 QAnon news posts.

using the original 26 topics and associated keywords without manual validation (see details in Suppl. S2.3). (iii) To ensure that our findings are not influenced by the topic granularity, we recalculate topic diversity at the theme level and repeat our analysis (see details in Suppl. S2.4). (iv) We further balance the samples between true and false news items through propensity score matching and reduce potential unobserved confounding factors (see details in Suppl. S2.5). (v) We extend conspiracy theories by including QAnon and repeat our analysis (see details in Suppl. S2.6).⁷ (vi) Given that news posts can also debunk linked false news items, we exclude potential debunking posts and repeat our analysis (see details in Suppl. S2.7). (vii) We use an alternative topic modeling method (BERTopic) to cluster topics again and repeat our analysis (see details in Suppl. S2.8). All the results are robust and consistently support our main findings.

2.4.5 Summary of Main Findings

Our analysis yields several key findings: (i) False news, especially when accompanied by conspiracy theories, exhibits higher topic diversity compared to true news. This discrepancy increases from source news claims to news posts on $\mathbb{X}$. (ii) At news source sharer level, neither the news lifetime nor the number of news posts is significantly associated with topic diversity. However, false news has a longer lifetime and receives more posts compared to true news. Additionally, the integration of conspiracy theories is associated with a longer lifetime for false news but has no significant association with the number of false news posts. (iii) At post viewer level, posts with higher topic diversity receive more reposts, likes, and replies, while this association is not stronger in false news posts. (iv) The integration of conspiracy theories is linked to more engagement with COVID-19 misinformation at post viewer level. False news posts that contain conspiracy theories, on average, receive 40.8% more reposts, 45.2% more likes, and 44.1% more replies compared to false news posts without conspiracy theories.

2.5 Discussion

2.5.1 Relevance

The spread of misinformation surrounding health crises is a serious public concern, intersecting multiple domains in society. Here, we present an empirical analysis of 7273 fact-checked source news claims and their associated posts on $\mathbb{X}$ to explore how topic diversity and conspiracy theories influence engagement with multi-topic misinformation during the COVID-19 pandemic. Understanding the characteristics of misinformation and their effects on user engagement is a crucial and growing area of research for developing effective countermeasures and mitigating the spread of misinformation [1, 10, 14, 33, 74]. Our study makes two main contributions: (i) We analyze the characteristics of misinformation and its spread from the perspectives of topic diversity and conspiracy theories; (ii) We examine how these characteristics shape engagement with misinformation at two distinct user levels, *i.e.*, news source sharer level and post viewer level. Specifically, we investigate online user engagement from the shares of source news items by news source sharers to the subsequent engagement with news posts by post viewers on $\mathbb{X}$.

(*i.e.*, reposts, likes, and replies). Our findings offer valuable insights for further research and practical strategies for combating the spread of online misinformation during health crises and beyond.

2.5.2 Research Implications

Heightened topic diversity in misinformation/conspiracy theories. The observed high topic diversity in COVID-19 misinformation suggests that the interconnection of multiple topics becomes a characteristic of misinformation narratives during health crises. Here, the topic diversity reflects the interconnectedness of various topics within a single piece of misinformation, potentially enhancing its novelty. Different from our study, Vosoughi et al. [14] investigated the topic similarity between users' reposted true (or false) posts and their previously viewed posts. They found that false posts are more dissimilar from previously viewed posts compared to true posts. However, given users' topic interests and recommendation algorithms' tendency to promote similar content, users potentially consume more content that aligns with their interests [88, 89]. These seemingly conflicting findings could be partially reconciled by our insights on topic diversity: misinformation may strategically link familiar topics together with diverse, novel topics, thereby drawing users' attention to new information and increasing their susceptibility. Particularly, our finding that misinformation containing conspiracy theories exhibits the strongest topic diversity underscores this deliberate strategy in conspiracy narratives. Moreover, previous research reveals coordinated efforts among misinformation websites from multiple domains to hyperlink to each other, further supporting our interpretation [42].

Additionally, we find that the topic diversity in COVID-19 misinformation significantly increases from its source claims to subsequent online posts. This mutation/evolution may contribute to the adaptability of misinformation, allowing it to resonate with diverse audiences and contexts [134]. For instance, QAnon is a far-right political conspiracy theory that originated in 2017. It became more diverse during COVID-19, evolving and expanding beyond political themes to incorporate pandemic-related topics [132]. Theoretically, within the context of multi-topic misinformation, individuals tend to adaptively connect related information and make associative inferences to reduce uncertainty and gain control in a way that is congruent to their prior knowledge [37, 90]. Understanding this cognitive process is essential for developing targeted interventions aimed at reducing susceptibility to multi-topic misinformation, especially when accompanied by conspiracy theories. Taken together, our findings underscore the importance of considering the multi-topic nature of misinformation in further research.

Engagement patterns at news source sharer level and post viewer level. During health crises marked by high uncertainty, novel information can update users' understanding and knowledge of the current state, helping them manage anxiety and fostering their engagement intentions [8, 10, 14, 83]. Furthermore, sharing novel posts can make users feel that they possess rare and important information that others may lack, thereby boosting their self-esteem [10, 135]. In this light, our findings suggest that the high topic diversity and the incorporation of conspiracy theories in misinformation can enhance its perceived novelty, thereby driving engagement. However, through further examining how topic diversity and conspiracy theories shape engagement at news source sharer level

and post viewer level, we find distinct engagement patterns at the two user levels.

Engagement patterns at news source sharer level: At news source sharer level, the observed longer lifetime and higher post count on $\mathbb{X}$ for false news items, compared to true news items, indicate the persistence and widespread reach of misinformation on social media during COVID-19. The persistence of false news items provides them with more opportunities to accumulate engagement on social media, potentially reaching a broader audience. Moreover, the lifetime and post count appear to be largely attributed to the veracity of news claims, as opposed to other variables including *Conspiracy* and *Diversity*. Given this, it is plausible that the prolonged lifetime and high post count for false news items mainly result from deliberately planned dissemination strategies by misinformation providers, involving means such as social bots and monetization [86, 95, 123]. This perspective also aligns with the prior finding that social bots amplify low-credibility content during the initial spreading stages [136]. Consequently, the role of individuals' self-motivated sharing behaviors in the persistence and diffusion of COVID-19 misinformation is limited at news source sharer level. However, massive posts repeating the same false claims with adaptability – *i.e.*, through increased topic diversity – can still heighten individuals' susceptibility and beliefs. In this regard, social media platforms should take measures to restrict repeated posts linking to the same false news items and their mutations.

Engagement patterns at post viewer level: As opposed to engagement at news source sharer level, engagement at post viewer level (*i.e.*, reposts, likes, and replies) on $\mathbb{X}$ might be largely driven by organic interactions from human users rather than social bots during COVID-19 [86]. Accordingly, we find that topic diversity and conspiracy theories become significant in driving engagement with misinformation at post viewer level in terms of reposts, likes, and replies. To efficiently reduce human users' engagement with misinformation at this level, social media platforms need to prioritize targeting misinformation that exhibits high topic diversity and incorporates conspiracy theories.

Additionally, our analysis reveals that replies differ from reposts and likes in their engagement with misinformation. Specifically, excluding the effects of topic diversity, conspiracy theories, and other control variables, false news posts receive more replies but fewer reposts and likes than true news posts. Distinct from reposts and likes, replies can not only support but also debunk false news posts. Given this, the higher number of replies for false news posts than true news posts can be triggered by debunking misinformation [137]. Our supplementary analysis of quotes shows that, although not statistically significant, false news posts also have a trend to receive more quotes than true news posts (see Suppl. S2.9 for full estimation results). This finding suggests that quotes serve a similar function as replies.

Taken together, our findings suggest that engagement with misinformation might be primarily driven by social bots at news source sharer level, while human users dominate engagement at post viewer level. This distinction highlights the need to examine misinformation engagement and develop intervention strategies at a multi-level framework.

2.5.3 Practical Implications

Implementing cross-platform fact-checking. Currently, social media providers are actively implementing measures to control the spread of misinformation on their platforms, including third-party professional fact-checking by experts and community-based fact-checking driven by crowds [1, 74]. However, these content moderation policies/features primarily function within individual platforms. Our findings suggest that misinformation from external sources, even if fact-checked on one platform, can still enter another platform through a large volume of posts. To prevent this spillover of misinformation across platforms, coordinated efforts among social media platforms, such as sharing fact-checking information, are needed.

Expanding fact-checks to misinformation with same links. Given that misinformation can be repeatedly shared on social media platforms through the same source links, an effective strategy is to directly moderate posts that contain previous fact-checked misinformation links. For example, X has rolled out a Community Notes feature that allows users to write fact-checking notes as additional context displayed on the corresponding misleading posts.⁸ This community-based fact-checking approach has proven effective in reducing the spread of misleading posts [76]. A recent update to Community Notes on X ensures that helpful notes that indicate posts' media elements, such as images or videos, are misleading will be visible on all posts containing the same media. Similarly, fact-checking notes about misinformation links could be expanded to cover all posts that contain the same links.

8: More details are available at <https://communitynotes.x.com/guide/en/about/introduction>

Empowering logical reasoning and critical thinking for social media users. To reduce users' susceptibility to misinformation with high topic diversity and promote rational information processing, tools that can enhance users' logical reasoning and critical thinking abilities should be adopted. The rise of Large Language Models (LLMs) presents a promising opportunity to help users conduct logical reasoning and identify potentially conflicting or fictional information. Although current LLMs may still have limitations in reasoning tasks, particularly for misinformation identification [138], this approach represents a promising direction as LLMs continue to improve [139].

2.5.4 Limitations and Further Research

Our study is subject to several limitations. First, our analysis is based on real-world social media data and, as is typical with observational studies, is limited to reporting associations between variables rather than causal claims. Future research could benefit from experimental designs and user studies to establish causal relationships and explore heterogeneity in the effects of topic diversity and conspiracy theories on engagement across different user groups.

Second, our findings on engagement can be influenced by social media recommendation algorithms and potential soft moderation implemented on misleading posts. Specifically, while our study mainly focuses on user-driven engagement at two levels and discusses the potential role of social bots, social media algorithms can selectively recommend posts and boost their visibility, potentially influencing user engagement and adding complexity to our results. Given that posts linking to the same news items can be treated similarly by recommendation algorithms or have similar audiences, the incorporation of news-specific random

effects, *i.e.*, random intercepts, in our models helps mitigate this concern. However, the precise role of recommendation algorithms warrants further exploration in future research.

Additionally, due to the absence of information indicating whether and when soft moderation was applied to certain misleading posts, we cannot control for the potential effect of soft moderation on subsequent engagement. Previous research has shown that soft moderation interventions – *e.g.*, attaching warning tags – adopted by X during COVID-19 can reduce belief in misinformation and thus lower the spread/visibility of misleading posts [140]. Moreover, a recent study demonstrates that community-based fact-checking approach is effective in reducing the spread of misleading posts on X [76]. With the potential impact of such interventions in reducing engagement, we still find that false news has a longer lifetime than true news. This indicates that our results are not overestimated. On the other hand, prior research suggests that warning labels can also prompt interactions aimed at debunking false claims [137]. To address this concern, we exclude potential debunking posts based on debunk-specific keywords and conduct our analysis again. Our results are robust and consistently support our main findings.

Furthermore, the absence of detailed engagement timelines at post viewer level in our dataset hinders a more granular analysis of how topic diversity and conspiracy theories influence engagement with misinformation over time on social media. This limitation highlights an important area for further examination.

Finally, our analysis is based on a single curated public dataset that includes social context on one platform – X [124]. Although this dataset offers unique features, such as multilingual content, it may introduce potential selection bias during data collection. To the best of our knowledge, no other publicly available datasets are suitable for addressing our research questions. Moreover, current restrictions on social media platform APIs make it increasingly challenging for researchers to access comprehensive data. We advocate for greater openness from social media platforms to facilitate academic research and encourage further studies to validate and generalize our findings across broader datasets and other social media platforms.

2.6 Conclusion

The spread of health misinformation on social media, especially during health crises such as COVID-19, poses unprecedented threats to our society. It is crucial to understand the factors and mechanisms that contribute to the engagement with health misinformation on social media. Here, we provide a comprehensive analysis of how topic diversity and conspiracy theories shape engagement with COVID-19 misinformation from news source sharer level to post viewer level. We find that high topic diversity and the integration of conspiracy theories are characteristics of COVID-19 misinformation. Topic diversity has no significant association with the engagement at news source sharer level. However, topic diversity and conspiracy theories can significantly enhance engagement with COVID-19 misinformation at post viewer level. Our findings provide insights into understanding the engagement with health misinformation and highlight the engagement patterns for news source sharers and post viewers regarding topic diversity and conspiracy theories. These insights

are valuable for developing targeted interventions at both user levels during health crises.

Fact-Checking Selection in Third-Party Organizations

3

Abstract: Political elites play an important role in the proliferation of online misinformation. However, an understanding of how fact-checking platforms pick up politicized misinformation for fact-checking is still in its infancy. Here, we conduct an empirical analysis of mentions of U.S. political elites within fact-checked statements. For this purpose, we collect a comprehensive dataset consisting of 35,014 true and false statements that have been fact-checked by two major fact-checking organizations (Snopes, PolitiFact) in the U.S. between 2008 and 2023, *i.e.*, within an observation period of 15 years. Subsequently, we perform content analysis and explanatory regression modeling to analyze how veracity is linked to mentions of U.S. political elites in fact-checked statements. Our analysis yields the following main findings: (i) Fact-checked false statements are, on average, 20% more likely to mention political elites than true fact-checked statements. (ii) There is a partisan asymmetry such that fact-checked false statements are 88.1% more likely to mention Democrats, but 26.5% less likely to mention Republicans, compared to fact-checked true statements. (iii) Mentions of political elites in fact-checked false statements reach the highest level during the months preceding elections. (iv) Fact-checked false statements that mention political elites carry stronger other-condemning emotions and are more likely to be pro-Republican, compared to fact-checked true statements. In sum, our study offers new insights into understanding mentions of political elites in false statements on U.S. fact-checking platforms, and bridges important findings at the intersection between misinformation and politicization.

3.1	Introduction	41
3.2	Background	43
3.3	Research Questions . . .	43
3.4	Data and Methods	45
3.4.1	Data Source: Fact-Checked Statements . . .	45
3.4.2	Politicization	45
3.4.3	Other-Condemning Emotions	47
3.4.4	Explanatory Regression Models	47
3.5	Empirical Analysis	48
3.5.1	Mentions of Political Elites (RQ1 & RQ2)	49
3.5.2	Elections (RQ3)	51
3.5.3	Other-Condemning Emotions (RQ4)	52
3.5.4	Exploratory Analysis & Robustness Checks	53
3.6	Discussion	55
3.7	Conclusion	56

Publication

Y. Chuai, J. Zhao, N. Pröllochs, and G. Lenzini, 'Is fact-checking politically neutral? Asymmetries in how U.S. fact-checking organizations pick up false statements mentioning political elites,' in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 19, 2025, pp. 403–429.

3.1 Introduction

Online misinformation poses significant challenges to societies [5]. Although the proportion of misinformation in (most) people's information diet tends to be small [30, 31, 33], detrimental effects of misinformation have been repeatedly observed across various events such as, for example, during elections and crisis situations [17, 18, 21, 32, 141]. Hence, countering the spread of misinformation has become an urgent priority, with extensive research efforts dedicated to analysis of its online diffusion and detection [1, 10, 14, 34, 74, 76, 96, 142]. However, despite these efforts, reducing people's susceptibility to misinformation remains challenging [1].

Recent research suggests that political elites (*i.e.*, politicians, political figures) may play a key role in the proliferation of online misinformation [17, 143]. Over the past 30 years, political figures have been increasingly mentioned in media representations with the goal of politicizing social

issues [144]. The strategy of mentioning politicians in news stories can drive engagement [102, 145] and help to foster specific political interests, ideologies, or agendas [13, 121, 144]. At the same time, politicization contributes to partisan and affective polarization, characterized by growing disparities in viewpoints between adherents of opposing political parties [146–148]. For instance, misinformation mentioning politicians has been shown to fragment online political debates and intensify divisions in public support for policies related to COVID-19 and climate change [143, 144, 149]. In light of a substantial body of research highlighting the influence of political elites on individuals' attitudes and behaviors [3, 17], the politicization of misinformation has become a pressing concern that requires further scrutiny.

Fact-checking organizations such as PolitiFact (politifact.com) and Snopes (snopes.com) are vital for identifying and debunking politicized misinformation. Their fact-checking assessments are supposed to help users to identify misleading statements in order to curb their spread (*e.g.*, on social media) [150]. Previous studies suggest that fact-checking can indeed improve misinformation discernment and reduce sharing intentions for misinformation [30, 44]. In practice, however, limited resources often force fact-checkers to prioritize statements for fact-checking [151]. Although their assessments are generally considered accurate [14, 152], the necessity of choosing specific claims to fact-check can invite criticism due to the subjective nature of the decision [153]. Critics argue that the selection of claims can reflect political preferences, potentially skewing results in favor of certain political identities (typically towards the political left) [154]. This is problematic as the perception of bias and subjectivity can erode public trust in fact-checking organizations. If people believe that fact-checkers are not impartial, they may be less likely to trust their findings, regardless of the evidence presented [56]. Despite these vocal concerns, empirical evidence on how fact-checking organizations pick up politicized misinformation for fact-checking is still missing.

Research goal: In this paper, we conduct an empirical analysis of mentions of political elites within fact-checked true and false statements. To this end, we collect 35,014 true and false statements that have been fact-checked by major U.S. fact-checking organizations, namely, PolitiFact (politifact.com) and Snopes (snopes.com), between 2008 and 2023, *i.e.*, within an observation period of 15 years. These two fact-checking organizations are among the largest and most established of their kind [14, 17, 21, 99]. Additionally, we construct a database of U.S. politicians, containing 3703 political elites (Republicans and Democrats) to measure politicization within the fact-checked statements. Subsequently, we estimate explanatory regression models to analyze how veracity is linked to mentions of U.S. political elites in fact-checked statements. We also study how this link varies between election cycles and across topics, and perform content analysis to explore the role of other-condemning emotions in the politicization of misinformation.

Contributions: Our work provides empirical evidence on how fact-checking organizations pick up politicized misinformation for fact-checking. We find that: (i) fact-checked false statements are, on average, 20% more likely to mention political elites than true fact-checked statements. (ii) There is a partisan asymmetry such that false fact-checked statements are 88.1% more likely to mention Democrats, but 26.5% less likely to mention Republicans, compared to true fact-checked statements. (iii) Mentions of political elites in fact-checked false statements reach the highest level during the months preceding elections. (iv) Fact-checked false statements that mention political elites carry stronger

other-condemning emotions and are more likely to be pro-Republican. These findings offer new insights into how major fact-checking platforms in the U.S. select their targets to fact-check, and bridge earlier work at the intersection between misinformation and politicization.

3.2 Background

Online misinformation is a pressing societal problem that platforms, policymakers, and researchers continue to grapple with. The reason is that there are serious concerns that misinformation on social media is damaging societies and democratic institutions [5, 33, 155]. The intersection of misinformation and political elites has garnered increased attention lately [17, 102, 143]. There has been a surge of news stories featuring political figures to politicize social issues. Politicization can drive engagement and help to foster political agendas [13, 102, 121, 144, 145]. Understanding politicization is important as it can contribute to partisan and affective polarization, which, in turn, strongly affects how people evaluate information [103, 143–145]. Existing works show that politicized misinformation can have profound impacts on voter attitudes and other offline behaviors [3, 17, 121, 144, 145]. For instance, the politicization of the (seemingly unpolitical) topic of climate change has contributed to a polarized environment that fosters climate change denial and misinformation [144]. Likewise, during elections, partisans have been observed to selectively share (oftentimes false) statements that express animosity toward political opponents (*i.e.*, members with the opposite political identity) [145]. Overall, the growing politicization is a pressing concern that has the potential to amplify the negative effects of misinformation on modern societies.

3.3 Research Questions

Previous studies indicate that political elites play a key role in the proliferation of misinformation. Yet, an understanding of how fact-checking platforms pick up politicized misinformation for fact-checking is still in its infancy. To address this gap, we propose four research questions (RQs1–4), each of which aims to extract new insights at the intersection of fact-checking and political elites.

Politicization (RQ1): Fact-checks carried out by fact-checking organizations are supposed to identify misinformation in order to inform the public and help curb its spread. However, the manual fact-checking process is both labor-intensive and time-consuming, and fact-checking organizations operate with limited resources [156]. This inherent limitation constrains the scale of fact-checking efforts and forces fact-checkers to prioritize statements based on their potential harm and urgency. Fragmentation and likelihood of spread are two main criteria for fact-checking prioritization [151]. Politicized misinformation is particularly prone to go viral, incites partisan and affective polarization, and poses significant threats to societies [17, 102, 143]. Here, mentioning political elites is a primary strategy to politicize misinformation stories [13, 121, 144]. Given this, fact-checking platforms may prioritize politicized statements mentioning political elites for fact-checking. However, empirical evidence on how the mentioning of political elites differs between fact-checked true

and false stories is missing. Accordingly, our first research question (RQ1) states:

***RQ1:** Are fact-checked false statements more likely to mention political elites than fact-checked true statements?*

Partisan asymmetry (RQ2): Politicization typically has a distinct partisan leaning, favoring either conservative (Republican, right-leaning) or liberal (Democratic, left-leaning) opinions [99]. Two factors lead us to anticipate that there may be a partisan asymmetry regarding true and false statements picked up for fact-checking by professional fact-checking organizations. First, there is ample evidence that the political left and right use the internet differently and right-leaning sources gain more visibility on social media, a phenomenon known as *ideological asymmetry* [157, 158]. Further, adherents of the right have been observed to be more likely to believe and share misinformation than adherents of the left [17, 99, 104, 145, 157]. Second, fact-checking organizations may more frequently target one political side or another. Anecdotal and small-scale empirical evidence suggest that fact-checkers in the U.S. target misinformation from Republicans more often than misinformation from Democrats [159]. Motivated by these two factors, our second research question (RQ2) is to gauge whether mentions of Democrats and Republicans vary across fact-checked true and false statements.

***RQ2:** Is there a partisan asymmetry in mentions of political elites across fact-checked true and false statements?*

Elections (RQ3): Elections are important political events in democracies, serving as focal points for political engagement and discourse. The campaigning activities of political candidates inherently generate a large number of (true and false) statements, which can affect voters' attitudes and behaviors [16, 18]. In response to the increased volume, professional fact-checking organizations may allocate more time and resources than usual to scrutinize politicized statements [151, 152]. Consequently, the level of politicization in fact-checked statements may rise as elections approach. Therefore, our third research question (RQ3) is:

***RQ3:** Are political elites more likely to be mentioned in fact-checked statements as elections draw closer?*

Other-condemning emotions (RQ4): A frequently employed strategy to politicize misinformation is to harness animosity toward political elites of the political out-group [32, 105]. Out-group animosity is an effective strategy for expressing one's partisan identity and generating engaging online content. For instance, it has been found that out-party animosity exerts a substantial influence on the spread of political misinformation when contrasted with in-party warmth [105]. Animosity toward the out-group is frequently expressed through the use of other-condemning emotions (*i.e.*, contempt, anger, and disgust), a subset of the moral emotions families [77, 81, 160]. Other-condemning emotions are reactions to the social behavior of others, involving a negative judgement or disapproval [81, 160]. Studying the role of other-condemning emotions thus allows one to better understand the context of politicization, *i.e.*, whether authors of fact-checked statements harness animosity toward the political out-groups. Accordingly, our fourth research question (RQ4) analyzes whether fact-checked statements that mention political elites contain more other-condemning emotions compared to fact-checked statements that do not mention political elites.

Table 3.1: Examples of fact-checked true and false statements. Mentions of political elites are highlighted in bold in the statements. The affiliated parties of mentioned politicians and other-condemning emotions in the statements are included.

Date	Statement	Affiliated Party	Other-Condemning	Veracity
2022-02-25	Gov. Tony Evers has “only gotten one-third of the money meant for COVID relief out the door. He is sitting on \$930 million in ARPA funds left unspent. In fact, he still has CARES Act money from two years ago.”	Democratic	0.069	True
2018-06-25	\$1 billion—that’s how much Bruce Rauner has wasted with his budget crisis.	Republican	0.750	False
2016-10-25	Wikileaks also shows how John Podesta rigged the polls by oversampling Democrats, a voter suppression technique.	Democratic	0.667	False
2014-10-23	Six convictions are connected to an allegation by prosecutors “that Gov. Scott Walker is at the center of a criminal scheme.”	Republican	0.571	False

RQ4: Do fact-checked statements that mention political elites contain more other-condemning emotions compared to fact-checked statements that do not mention political elites?

3.4 Data and Methods

3.4.1 Data Source: Fact-Checked Statements

For our analysis, we collect a large-scale dataset consisting of true and false statements that have been fact-checked between 2008 and 2023 (*i.e.*, within an observation period of 15 years) by two major fact-checking organizations, namely, PolitiFact (politifact.com) and Snopes (snopes.com). We scrape all fact-checked statements, including associated information such as verdicts, fact-checking dates, fact-checkers, and topic tags from these websites (see examples in Table 3.1). As a result, our dataset comprises a total of 35,014 fact-checked statements, with 22,561 statements originating from PolitiFact and 12,453 statements originating from Snopes. The fact-checks have been carried out by 608 different fact-checkers. Given that PolitiFact and Snopes have different scales for rating the veracities of the statements, we follow earlier research [14, 161] and classify the ratings into distinct true or false categories.¹ Specifically, PolitiFact provides six truthfulness ratings, namely, true, mostly true, half true, mostly false, false, and pants on fire. We consider half true, mostly false, false, and pants on fire as false ratings; and the other categories as true ratings. For Snopes, we consider mixture, mostly false, false, miscaptioned, misattributed, scam, and fake as false ratings. True, mostly true, correct attribution, legit, and recall are taken as true ratings.

1: Analogous to previous research [161], we consider fact-checked statements with mixed veracity (*e.g.*, half-true and mixture) as false ratings. The results in our later analysis are qualitatively identical if we exclude statements with mixed veracity (see Suppl. S3.6.2).

In addition to the fact-checking verdicts, we clean the fine-grained topic tags provided by the fact-checking organizations and integrate them into our dataset. Specifically, there are in total 202 topic tags categorized by the two organizations. The topic tags include many specific social issues, such as, for example, Coronavirus, Guns, and Immigration. Summary statistics of the dataset and most frequent topics are reported in Tables S3.1 and S3.2.

3.4.2 Politicization

We examine the politicization of fact-checked true and false statements based on the mentions of political elites (see Table 3.1), which is the core of

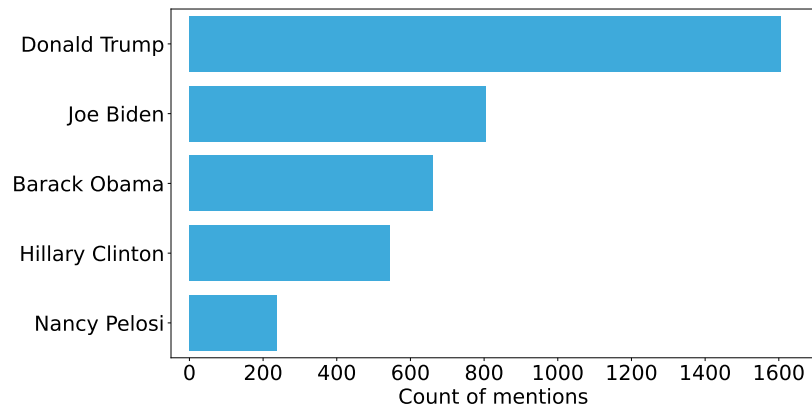


Figure 3.1: The five most frequently mentioned politicians in fact-checked statements

2: <https://web.archive.org/web/20230723193207/https://www.politifact.com/personalities/>
 3: <https://www.congress.gov/members>

most conceptualizations of politicization [121, 144]. For this purpose, we gather a comprehensive list of U.S. political elites based on two sources: (i) we collect 4919 public figure names from the People section on the web page of PolitiFact.² (ii) We additionally collect the names of all 1254 members of the U.S. Congress since 2008 from the official Congress website.³ We merge both data sources together and only consider the political elites affiliated with either the Republican or the Democratic party. The resulting list includes the names of 3703 political elites, out of which 1938 are Republicans and 1765 are Democrats. Note that the list of political elites used in this paper is substantially more comprehensive than the lists used in earlier research [17, 105, 121]. This allows us to maximize the matches with political elites mentioned in the fact-checked statements.

We use the list of political elites to perform string matching of the fact-checked statements. For this purpose, we first preprocess the texts by removing non-alphabetic and non-numerical characters. Subsequently, we use the names of political elites as keywords to match the cleaned texts. Statements are considered politicized when they mention one or more political elites. Out of 35,014 fact-checked statements, 9308 (26.6%) are categorized as politicized. Among these, 7351 (79%) statements mention only one politician, and 1957 (21%) statements mention two or more politicians. Notably, only 674 (7.2%) statements mention Democrats and Republicans at the same time. This suggests that most of the politicized statements have a specific focus on either Democrats or Republicans.

Fig. 3.1 shows that Donald Trump, Joe Biden, Barack Obama, Hillary Clinton, and Nancy Pelosi are the five most frequently mentioned politicians in the fact-checked statements. The number of fact-checked statements that mention one of these five politicians is 3670, which accounts for 39.4% of all politicized statements.

Manual validation: To evaluate the performance of the string matching approach, we manually assess 200 fact-checked statements randomly selected from our dataset. The weighted average scores of precision and recall are 91% and 89%, respectively, which indicates a high performance of the string-matching approach. As a further check, we evaluate all politicized statements by additionally extracting fact-checking comments, *i.e.*, the textual verdicts of fact-checkers, for the corresponding statements (see Suppl. S3.2). Here, we find an almost perfect precision score of 99.6%.

Partisan scores: Additionally, each political elite has a distinct political affiliation, namely, Republican or Democratic. We code the statements according to the political affiliations of the political elites mentioned

in the textual content. If the political elite is Democrat, the partisan score is -1 , and if the political elite is Republican, the partisan score is 1 . If the statements contain the names of more than one political elite, we sum the partisan scores. For example, if a statement mentions one Republican and one Democrat, then the statement has a partisan score of 0 . A positive score (> 0) indicates that the statement predominantly mentions Republicans, whereas a negative score (< 0) signifies a predominant mentioning of Democrats. Out of all fact-checked statements, 4403 statements predominantly mention Republicans, and 4455 statements predominantly mention Democrats.

3.4.3 Other-Condemning Emotions

Other-condemning emotions are a subset of moral emotions and comprise the emotions contempt, anger, and disgust [81, 160]. We use a dictionary-based approach to computationally measure the expression of other-condemning emotions. Specifically, we employ the NRC emotion lexicon to match emotion words in the cleaned text of the fact-checked statements. The NRC emotion lexicon is a widely used tool in previous research [10, 14, 34, 81, 162]. It includes 14,154 unique English words and their associations with eight basic emotions [163]. Additionally, the eight basic emotions, namely, anger, anticipation, disgust, fear, joy, sadness, surprise, and trust, can be combined into 24 complex emotional dyads [128]. For example, contempt is composed of anger and disgust. In line with the prior work that has validated the effectiveness of using NRC emotion lexicon to measure other-condemning emotions [81], we calculate the scores of the eight basic emotions and the 24 derived emotions based on the frequency of words belonging to each emotion category. Subsequently, the score of other-condemning emotions is calculated as the sum of the scores for contempt, anger, and disgust. The mean score for other-condemning emotions in the fact-checked statements is 0.096 (see Table S3.1 for details). This indicates that, on average, the three other-condemning emotions account for 9.6% of all emotion components from Plutchik's emotion model.

3.4.4 Explanatory Regression Models

In this study, we employ explanatory regression modeling to answer our research questions. The variables included in our models are as follows (summary statistics are in Table S3.1):

- **Politician**: A dummy variable indicating that the fact-checked statement explicitly mentions politician(s) ($=1$; otherwise 0).
- **Republican**: A dummy variable indicating that the fact-checked statement predominantly mentions Republicans ($=1$; otherwise 0).
- **Democrat**: A dummy variable indicating that the fact-checked statement predominantly mentions Democrats ($=1$; otherwise 0).
- **Falsehood**: A dummy variable indicating that the fact-checked statement is labeled as false ($=1$; otherwise 0).
- **MTE** (Months to Elections): A continuous variable measuring the time interval (in months) between the date of each fact-checked statement and the next upcoming election. Here, we consider 2-year election cycles including both U.S. presidential and midterm elections.
- **Words**: The number of words in the cleaned text of each fact-checked statement. This (control) variable is commonly used to account for the content richness of the statements [74].

- **OtherCondemning**: A continuous variable measuring other-condemning emotions in fact-checked statements.

To address research questions RQ1, RQ2, and RQ3, we treat *Politician*, *Republican*, and *Democrat* as dependent variables. The explanatory variables are *Falsehood*, *MTE*, and *Words*. Note that some fact-checkers might be more focused on politics and, thus, may prefer to check statements that mention political elites. Hence, we control for such heterogeneity among fact-checkers through the use of fact-checker-specific random effects (*i.e.*, random intercepts). Formally, we specify the following three logistic regression models (one for each of the three dependent variables):

$$\begin{aligned} \text{logit}(y) = & \beta_0 + \beta_1 \text{Falsehood}_i + \beta_2 \text{MTE}_i + \beta_3 \text{Falsehood}_i \times \text{MTE}_i \\ & + \beta_4 \text{Words}_i + u_{\text{year}} + u_{\text{factchecker}}, \end{aligned} \quad (3.1)$$

where y is the binary dependent variable (*i.e.*, *Politician*, *Republican*, or *Democrat*), β_0 is the intercept, β_1 to β_4 are the coefficients of the explanatory variables, u_{year} are yearly fixed effects, and $u_{\text{factchecker}}$ are random effects specific to the fact-checkers. Notably, we also include an interaction term $\text{Falsehood}_i \times \text{MTE}_i$, which allows us to analyze whether the level of politicization in fact-checked false statements differs from that in fact-checked true statements during election periods (RQ3).

To address RQ4, we consider *OtherCondemning* as the dependent variable, while treating *Politician* as an independent variable. The rationale is that the politicization of statements may be linked to the expression of other-condemning emotions [105, 160]. Consequently, we specify the following linear regression model with fact-checker-specific random effects and yearly fixed effects:

$$\begin{aligned} \text{OtherCondemning} = & \beta_0 + \beta_1 \text{Politician}_i + \beta_2 \text{Falsehood}_i + \beta_3 \text{MTE}_i \\ & + \beta_4 \text{Words}_i + u_{\text{year}} + u_{\text{factchecker}}. \end{aligned} \quad (3.2)$$

Note that in our later analysis, we also consider the interactions of *Politician* with both *Falsehood* and *MTE*. This allows us to investigate potential variations concerning the politicization of true and false fact-checked statements and how the effects vary during elections.

3.5 Empirical Analysis

Our dataset consists of 35,014 true and false statements that have been fact-checked by major U.S. fact-checking organizations between January 2008 and July 2023. The dataset thus spans an observation period of approximately 15 years, including eight U.S. Midterm and Presidential elections. The number of fact-checked false statements tends to increase when approaching elections, with the highest peak observed during the 2016 U.S. Presidential election (Fig. 3.2a). In contrast, the number of fact-checked true statements is relatively stable with a significant peak after the 2020 U.S. Presidential election. Fig. 3.2b shows that the mean of the variable Months to Elections (*MTE*) in fact-checked false statements (mean of 11.321) is smaller than that in fact-checked true statements (mean of 12.349). This implies that fact-checks for false statements are more concentrated close to elections than those for true statements. Furthermore, the mean of *MTE* in fact-checked statements mentioning political elites (mean of 10.031) is smaller than that in fact-checked statements without mentioning political elites (mean of 12.173). This

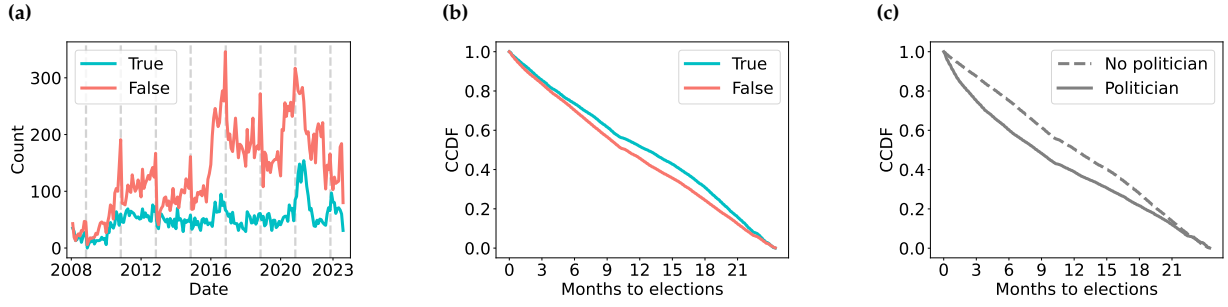


Figure 3.2: Dataset overview. (a) The monthly count of fact-checked true and false statements in our observation period between 2008 and 2023. The eight grey vertical dash lines correspond to eight U.S. election dates. (b) Complementary Cumulative Distribution Functions (CCDFs) for *MTE* (months to elections) in fact-checked true and false statements. The difference in distributions is statistically significant according to a KS-test ($KS = 0.073, p < 0.001$). (c) CCDFs for *MTE* in fact-checked statements w/ and w/o mentions of political elites. The difference in distributions is statistically significant according to a KS-test ($KS = 0.151, p < 0.001$).

suggests that as elections approach, fact-checkers are more likely to target politicized statements over non-politicized ones (Fig. 3.2c).

3.5.1 Mentions of Political Elites (RQ1 & RQ2)

We estimate logistic regression models to analyze how veracity ratings are linked to mentions of U.S. political elites (Republicans and Democrats) in the fact-checked statements (RQ1 & RQ2). The regression results are reported in Table 3.2. Columns (1), (2), and (3) correspond to the dependent variables *Politician*, *Republican*, and *Democrat*, respectively. As shown in Column (1), the coefficient estimate for *Falsehood* is statistically significant and positive ($coef. = 0.182, p < 0.001$), with an odds ratio (*o.r.*) of $e^{0.182} = 1.200$ (95% CI: [1.083, 1.330]). This implies that false statements are 20.0% more likely to mention political elites, as compared to true statements (RQ1).

Subsequently, we examine how the mentions of Republicans and Democrats differ between fact-checked true and false statements. For *Republican* [Column (2)], we find that the coefficient estimate for *Falsehood* is significantly negative ($coef. = -0.308, p < 0.001$; *o.r.* = 0.735, 95% CI: [0.648, 0.833]). This implies that false statements are 26.5% less likely to mention Republicans compared to true statements. Conversely, when examining *Democrat* [Column (3)], we find a statistically significant positive coefficient estimate for *Falsehood* ($coef. = 0.632, p < 0.001$; *o.r.* = 1.881, 95% CI: [1.630, 2.172]). This indicates that fact-checked false statements are 88.1% more likely to mention Democrats than fact-checked true statements. Additionally, the higher likelihood of mentioning Democrats in fact-checked false statements compared to fact-checked true statements is robust across topics (see Suppl. S3.3 for details). In summary, our findings imply that fact-checked false statements are more likely to mention Democrats and less likely to mention Republicans than fact-checked true statements (RQ2).

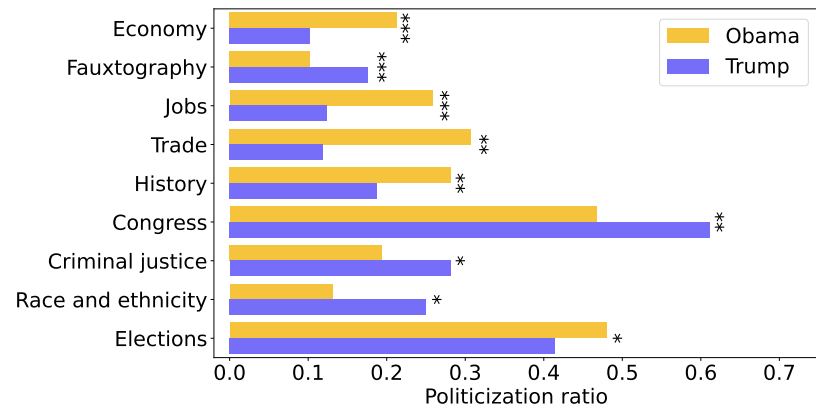
Analysis across incumbent U.S. presidents: The observation period covered by our dataset includes three full presidential cycles, *i.e.*, two presidential terms from Barack Obama (from 2008-11-04 through 2012-11-06 to 2016-11-08, Obama’s incumbency) and one presidential term from Donald Trump (from 2016-11-08 to 2020-11-03, Trump’s incumbency).⁴ Specifically, there are 14,515 fact-checked statements during Obama’s incumbency and 11,452 fact-checked statements during Trump’s incumbency, respectively. We incorporate an incumbency fixed effect (*Trump*)

4: Additionally, our dataset covers two years of the presidency of Joe Biden. To ensure comparability with other presidents’ incumbencies, we omit the period of Biden’s incumbency and analyze the temporal trend of mentions of political elites during the period from 2008 to 2020 and across Obama’s and Trump’s incumbencies.

Table 3.2: Estimation results for mixed logistic regression models predicting whether the fact-checked statements mention political elites of either political party [Column (1)], Republicans [Column (2)], or Democrats [Column (3)]. Fact-checker-specific random effects (REs) and yearly fixed effects (FEs) are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) Politician	(2) Republican	(3) Democrat
Falsehood	0.182*** (0.052)	-0.308*** (0.064)	0.632*** (0.073)
MTE	-0.037*** (0.004)	-0.033*** (0.005)	-0.033*** (0.006)
Falsehood $\times$ MTE	0.007 (0.004)	0.018*** (0.005)	-0.004 (0.006)
Words	0.022*** (0.002)	0.019*** (0.002)	0.014*** (0.002)
Intercept	-1.225*** (0.134)	-2.157*** (0.172)	-2.119*** (0.153)
Fact-checker REs	✓	✓	✓
Year FEs	✓	✓	✓
#Statements	35,014	35,014	35,014

Figure 3.3: The politicization ratios of the nine topics that have statistically significant differences during Obama's and Trump's incumbencies. The Chi-Square test is conducted for each topic group. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.



and its interaction with *Falsehood* into the regression models to examine whether there is a difference in mentioning political elites in fact-checked statements between Obama's incumbency and Trump's incumbency.

According to the estimation results [see Column (1) of Table S3.3], the coefficient estimate of *Trump* is significantly positive (*coef.* = 0.174, $p < 0.05$; *o.r.* = 1.190, 95% CI: [1.039, 1.363]). This indicates that fact-checked statements (both true and false) during Trump's incumbency, were 19% more likely to mention political elites (either Democrats or Republicans) compared to those during Obama's incumbency. However, when we only consider mentions of Republicans [see Column (2) of Table S3.3], the coefficient estimate of *Falsehood* $\times$ *Trump* is significantly negative (*coef.* = -0.258, $p < 0.01$; *o.r.* = 0.772, 95% CI: [0.651, 0.916]). This implies that fact-checked false statements during Trump's incumbency were 22.8% less likely to mention Republicans than those during Obama's incumbency. Overall, this indicates a more pronounced partisan asymmetry in fact-checking activity during Trump's incumbency compared to Obama's incumbency.

We further examine how the politicization of individual topics (*e.g.*, Economy, Race and Ethnicity, etc.) has changed across presidential incumbencies. To this end, we calculate a politicization ratio for each of the topics measuring the proportion of statements mentioning political elites relative to all fact-checked statements on the same topic. We find that the majority of topics have similar politicization ratios between Obama's incumbency and Trump's incumbency, with only nine topics

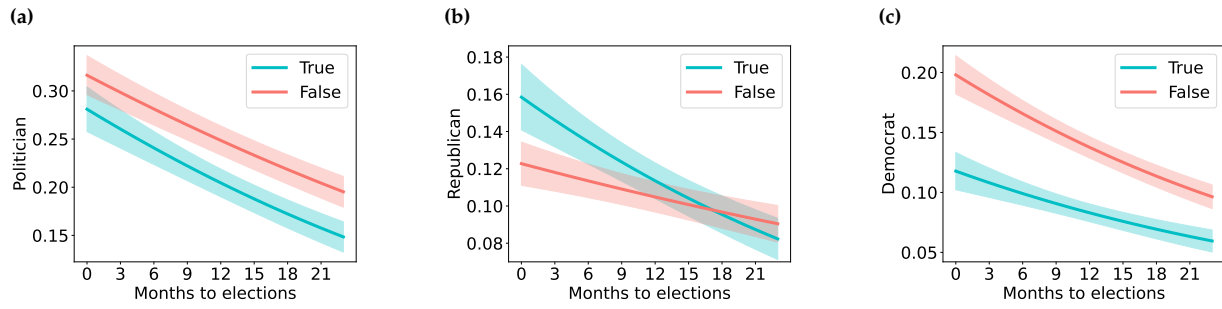


Figure 3.4: Predicted marginal means for the probabilities of mentions of (a) political elites of either political party, (b) Republicans, and (c) Democrats in fact-checked true and false statements. The error bands represent 95% Confidence Intervals (CIs).

having significant differences. Specifically, as shown in Fig. 3.3, the topics of “Economy,” “Jobs,” “Trade,” “History,” and “Elections” have higher politicization ratios during Obama’s incumbency, while the topics of “Fauxtography,” “Congress,” “Criminal justice,” and “Race and ethnicity” have higher politicization ratios during Trump’s incumbency (see Suppl. S3.3 for details).

3.5.2 Elections (RQ3)

To answer RQ3, we examine how the likelihood of mentions of political elites in fact-checked statements varies with regard to the time distance to U.S. elections. The coefficient estimates for the variable *MTE* (see Table 3.2) are consistently negative and statistically significant [*coef.* = -0.037 , $p < 0.001$, *o.r.* = 0.963 , 95% CI: $[0.956, 0.970]$ in Column (1); *coef.* = -0.033 , $p < 0.001$, *o.r.* = 0.967 , 95% CI: $[0.958, 0.976]$ in Column (2); and *coef.* = -0.033 , $p < 0.001$, *o.r.* = 0.968 , 95% CI: $[0.957, 0.978]$ in Column (3)]. This suggests that fact-checked statements are more likely to mention political elites (both Republicans and Democrats) as elections approach. Specifically, one month closer to elections corresponds to a 3.7% increase in the odds of mentioning political elites of either Republicans or Democrats in fact-checked statements. Furthermore, the coefficient estimate for the interaction *Falsehood* $\times$ *MTE* in Column (2) of Table 3.2 is significantly positive (*coef.* = 0.018 , $p < 0.001$; *o.r.* = 1.018 , 95% CI: $[1.008, 1.028]$). This means that, as time progresses away from elections, the likelihood of mentioning Republicans in fact-checked false statements decreases at a slower rate compared to that in fact-checked true statements. However, the coefficient estimate for *Falsehood* $\times$ *MTE* in Column (3) of Table 3.2 is not statistically significant (*coef.* = -0.004 , $p > 0.05$; *o.r.* = 0.996 , 95% CI: $[0.985, 1.007]$), which indicates that there is no corresponding significant difference in the changes in mentions of Democrats before elections between fact-checked true and false statements.

The predicted marginal means⁵ for fact-checked true and false statements following elections are visualized in Fig. 3.4. Consistent with our assessment of the coefficient estimates, fact-checked statements are more likely to mention politicians and especially Democrats, with the probability of mentioning political elites increasing in parallel for both true and false statements when elections approach (RQ3).

5: Within our logistic regression models, the predicted marginal means indicate the probability of mentioning political elites in fact-checked statements.

3.5.3 Other-Condemning Emotions (RQ4)

We now analyze whether fact-checked statements that mention political elites express more other-condemning emotions compared to fact-checked statements that do not mention political elites (RQ4). The regression results are reported in Table 3.3. We first examine a baseline model without interactions [Column (1)]. The coefficient estimate for *Politician* is significantly positive ($coef. = 0.009, p < 0.001$). Specifically, the predicted marginal means for other-condemning emotions in politicized and non-politicized statements are 0.102 (95% CI: [0.097, 0.106]) and 0.093 (95% CI: [0.090, 0.096]), respectively. This implies that fact-checked statements with political elites express 9.7% more other-condemning emotions compared to those without political elites (RQ4). Additionally, the significantly positive coefficient estimate for *Falsehood* suggests that fact-checked false statements contain more other-condemning emotions compared to fact-checked true statements ($coef. = 0.007, p < 0.01$). The predicted marginal means for other-condemning emotions in fact-checked true and false statements are 0.090 (95% CI: [0.086, 0.094]) and 0.097 (95% CI: [0.094, 0.100]), respectively. This implies that fact-checked false statements express 7.8% more other-condemning emotions compared to fact-checked true statements. Notably, the coefficient estimate for *MTE* is not statistically significant ($coef. = 0.000, p > 0.05$), indicating that the expression of other-condemning emotions is not significantly linked to time to elections.

We further examine the moderating effects of the variables *Falsehood* and *MTE* on *Politician*. As shown in Column (2) of Table 3.3, the coefficient estimate for the interaction *Politician* $\times$ *Falsehood* is not statistically significant ($coef. = -0.004, p > 0.05$), while the coefficient estimate for *Falsehood* is significantly positive ($coef. = 0.008, p < 0.01$). This suggests that fact-checked false statements carry more other-condemning emotions compared to fact-checked true statements, regardless of whether political elites are mentioned. In contrast, the coefficient estimate for *Politician* $\times$ *MTE* is significantly negative ($coef. = -0.001, p < 0.01$), whereas the coefficient estimate for *MTE* is not significant ($coef. = 0.000, p > 0.05$). This suggests that the expression of other-condemning emotions in fact-checked statements mentioning political elites decreases as the time period moves further away from elections.

Additionally, we include *Republican* and *Democrat* in the model [Column (3) of Table 3.3]. However, we find that the coefficient estimates for the two variables are both not statistically significant (each $p > 0.05$). This indicates that mentioning political elites, either Republican or Democrat, has a similarly positive association with other-condemning emotions within fact-checked statements.

Analysis across incumbent U.S. presidents: To analyze differences in other-condemning emotions of fact-checked statements between Obama's and Trump's incumbencies, we incorporate incumbency fixed effect (*Trump*) and its interactions with *Politician* and *Falsehood* into the base model. According to the estimation results [see Column (1) of Table S3.5], the coefficient estimate of *Trump* is significantly positive ($coef. = 0.014, p < 0.01$), which indicates that other-condemning emotions in fact-checked statements during Trump's incumbency are higher than that during Obama's incumbency. Moreover, the coefficient estimates of *Politician* $\times$ *Trump* ($coef. = -0.008, p > 0.05$) and *Falsehood* $\times$ *Trump* ($coef. = 0.005, p > 0.05$) are not statistically significant. This suggests that the high other-condemning emotions during Trump's incumbency are not

	(1) Baseline	(2) Interactions	(3) Partisanship
Politician	0.009*** (0.002)	0.020*** (0.006)	0.025* (0.010)
Republican			−0.005 (0.009)
Democrat			−0.004 (0.009)
Falsehood	0.007** (0.002)	0.008** (0.003)	0.008** (0.003)
Politician × Falsehood		−0.004 (0.005)	−0.004 (0.005)
MTE	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Politician × MTE		−0.001** (0.000)	−0.001** (0.000)
Words	0.001*** (0.000)	0.001*** (0.000)	0.001*** (0.000)
Intercept	0.099*** (0.009)	0.095*** (0.009)	0.095*** (0.009)
Fact-checker REs	✓	✓	✓
Year FEs	✓	✓	✓
#Statements	35,014	35,014	35,014

Table 3.3: Estimation results for mixed linear regression models predicting other-condemning emotions in the fact-checked statements. Columns (1)–(3) report the estimation results from the baseline model, model with interactions, and model with party indicators of (*Republican* and *Democrat*), respectively. Fact-checker-specific random effects and yearly fixed effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

specific to politicized or false statements but general to all fact-checked statements.

3.5.4 Exploratory Analysis & Robustness Checks

Partisan leanings of fact-checked statements: Our study suggests that fact-checked false statements are more likely to mention Democrats compared to fact-checked true statements. However, previous research indicates that widely shared fact-checked false statements tend to be pro-Republican (*i.e.*, benefiting the political right, right-leaning) [99, 164]. This dissonance may possibly be reconciled by fact-checked false statements being more likely to harness animosity toward Democrats compared to true statements. To test this notion, we further examine the use of other-condemning emotions toward political elites and explore the partisan leanings of fact-checked true and false statements. Here, we combine mentions of politicians (Republicans or Democrats) and other-condemning emotions together to define a new dependent variable *PartisanLeaning*. We code fact-checked statements as right-leaning (pro-Republican) if they utilize other-condemning emotions within fact-checked statements that predominantly mention Democrats; and vice versa. Formally, we define the dependent variable *PartisanLeaning* as $-PartisanScore \times OtherCondemning$, where *PartisanLeaning* > 0 means that the statement is right-leaning (pro-Republican), and *PartisanLeaning* < 0 means that the statement is left-leaning (pro-Democrat).

Subsequently, we estimate mixed linear regressions predicting partisan leanings in fact-checked true and false statements (see Suppl. S3.5 for details and full estimation results). For all fact-checked statements, the coefficient estimate of the explanatory variable *Falsehood* is significantly positive [*coef.* = 0.009, $p < 0.001$; see Column (1) of Table S3.6]. This

indicates that fact-checked false statements are more pro-Republican and use higher other-condemning emotions when mentioning Democrats, compared to true statements. The predicted marginal means for partisan leanings in true and false statements are -0.005 (95% CI: $[-0.007, -0.002]$) and 0.004 (95% CI: $[0.002, 0.006]$), respectively. This indicates that, on average, fact-checked false statements are pro-Republican, whereas fact-checked true statements are pro-Democrat.

Furthermore, PolitiFact provides author information for the fact-checked statements. This allows us to partially identify the political elites who authored the fact-checked statements based on our politician database (see Suppl. S3.5 for details). Specifically, there are 13,693 fact-checked statements from political elites, among which 5640 from Democrats and 8053 from Republicans. Given this, we conduct a subgroup analysis for fact-checked statements from PolitiFact and validate whether politicians express more other-condemning emotions in the fact-checked statements that mention political out-groups compared to non-partisan authors. According to the estimation results [see Column (3) of Table S3.6], fact-checked statements from Democrats are more pro-Democrat than those from non-Democrats ($coef. = -0.021, p < 0.001$). In contrast, fact-checked statements from Republicans are more pro-Republican than those from non-Republicans ($coef. = 0.012, p < 0.001$). Further details and full estimation results are provided in Suppl. S3.5.

Taken together, concordant with our hypothesis, we find that fact-checked false statements are indeed more pro-Republican than true fact-checked statements. Furthermore, fact-checked false statements from Republicans are more pro-Republican and express higher other-condemning emotions when mentioning Democrats compared to fact-checked true statements; and we observe the same pattern for fact-checked false statements from Democrats. These results strengthen the interpretation of our findings and suggest that fact-checked false statements from partisans are more likely to harness animosity toward political out-groups than fact-checked true statements.

Additional robustness checks: We conduct several checks to validate the robustness of our results: (1) We check for possible multicollinearity issues. The cross-correlations among the independent variables in our analysis are fairly small. Further, the variance inflation factors for the independent variables are all close to one; and, thus, well below the critical threshold of four (see details in Suppl. S3.6.1). Taken together, these findings indicate that multicollinearity is not an issue in our analysis. (2) We exclude fact-checked statements with mixed veracity ratings and repeat our analysis (see details in Suppl. S3.6.2). (3) We correct false positives in the manually checked politicized statements and repeat out analysis (see details in Suppl. S3.6.3). (4) We exclude fact-checked statements that contain mentions of both Republicans and Democrats and repeat out analysis (see details in Suppl. S3.6.4). (5) We include additional fixed effects for the fact-checking organizations and specific fact-checkers; see details in Suppl. S3.6.5. (6) We include author-level fixed effects, *i.e.*, we additionally control for the party affiliations of the authors of the fact-checked statements (see details in Suppl. S3.6.6). (7) We conduct propensity score matching to balance the samples in our dataset and repeat our analysis (see details in Suppl. S3.6.7). Our findings are consistently supported by all of these checks.

3.6 Discussion

Politicization plays a critical role in driving engagement with online misinformation [14, 17, 34, 99, 102, 105, 145]. As a remedy, fact-checking organizations (*e.g.*, PolitiFact, Snopes) are supposed to identify and debunk politicized misinformation to curb its spread. However, an understanding of how fact-checking platforms pick up politicized misinformation for fact-checking is largely missing. Here, we contribute to research at the intersection of misinformation and politicization by conducting a large-scale empirical analysis characterizing mentions of political elites within true and false statements that have been subject to fact-checking by major U.S. fact-checking platforms.

Implications: Our findings imply that major fact-checking organizations in the U.S. regularly fact-check politicized misinformation, especially during election months. However, there is a pronounced partisan asymmetry: false statements selected for fact-checking are significantly more likely to be pro-Republican, while the selected true statements are more likely to be pro-Democrat. This finding aligns with earlier conjectures that fact-checking organizations in the U.S. tend to fact-check the political right more than the political left [153, 154, 165]. Importantly, however, this observed asymmetry does not necessarily imply political bias on the part of the fact-checking platforms in their reporting of politicized false statements. Instead, it could also reflect the possibility that adherents of the political right produce a higher volume of (viral) misinformation that requires fact-checking. Previous research suggests that widely shared false political statements in the U.S. are more often pro-Republican [99, 164] and that the political right is more likely to hold misconceptions than the political left [17]. Disentangling these two factors likely requires a representative sample of *all* circulating misinformation. This presents an important – yet difficult – challenge for future research.

We further observe a pronounced role of out-group animosity in the politicization of misinformation. Partisans from both sides of the political spectrum frequently harness animosity toward political out-groups in their false statements. This out-group animosity is stronger than usual during election periods. Such a high prevalence of misinformation about political out-groups – at least within statements that undergo fact-checking – is worrying as it could imply that more in-group members believe and share this kind of misinformation [103, 105].

From a practical perspective, our study emphasizes the need for close monitoring of the activities and potential biases of major fact-checking organizations such as Snopes and PolitiFact. These platforms play a significant role in shaping public opinions and discourse in an era of widespread misinformation. Additionally, many research findings depend on the assessments provided by these organizations [14, 17, 21, 34, 81, 99, 105]. If fact-checking organizations exhibit bias – whether intentional or unintentional – it could skew public perception, undermine trust in research findings, and diminish confidence in fact-checking as an (objective) tool to combat misinformation. Hence, it is essential that fact-checking organizations not only conduct their work impartially but also maintain transparency about their practices (*e.g.*, by being open about their criteria for selecting claims). Overall, continuous oversight and transparency are vital for the integrity of the fact-checking process and ensuring that the public (and the academic community) can rely on their evaluations.

Limitations and future research: Our work has a number of limitations, which can fuel future research as follows. First, our analysis does not reveal whether the observed (political) asymmetries in fact-checking activities reflect political preferences/biases of the fact-checking organizations; or rather a result of adherents to the political left/right producing misinformation at different rates. Yet, disentangling this dual role is challenging as the population of misinformation is unknown, which provides an important challenge for further research. Second, we conduct our analysis based on a large sample of fact-checked statements collected from two fact-checking organizations: Snopes and PolitiFact. These are major fact-checking organizations in the U.S. and have been used as a main source of fact-checking verdicts in prior misinformation research. To the best of our knowledge, there are currently no other fact-checking sources that can provide comparable fact-checked statements over 15 years. Notwithstanding, to further generalize our findings, one could collect fact-checked statements from other fact-checking sources when available in the future. Third, this study primarily measures politicization through mentions of formal names of political elites, which might miss some other forms of politicization (*e.g.*, nicknames of politicians and non-official political activists). Nevertheless, focusing on mentions of formal names of political elites is a common choice in previous research and ensures that we do not overestimate politicization in false statements [144]. Finally, this paper focuses on statements in English and the context of U.S. politics. More data from other countries and languages need to be collected to foster a more comprehensive understanding of the link between politicization and misinformation across cultures.

3.7 Conclusion

Political elites are increasingly mentioned in the discussion of social issues and play a critical role in driving engagement with online misinformation. However, an understanding of how fact-checking platforms pick up and report false statements that mention political elites of either Republicans or Democrats remains underexplored. Here, we provide a comprehensive empirical analysis of mentions of political elites in true and false statements that have been subject to fact-checking by two major U.S. fact-checking organizations. We find that fact-checked false statements are more likely to mention political elites than fact-checked true statements, especially Democrats and during election months. Furthermore, we demonstrate a pronounced role of out-group animosity in politicized misinformation. Specifically, fact-checked false statements that mention political elites carry stronger other-condemning emotions and are more likely to be pro-Republican, while fact-checked true statements are more likely to be pro-Democrat. Taken together, our findings offer new insights into how fact-checking platforms in the U.S. fact-check politicized misinformation, and bridge important findings on the intersection between misinformation and politicization.

PART II: EFFECTIVENESS OF COMMUNITY-BASED FACT-CHECKING

Abstract: Developing interventions that successfully reduce engagement with misinformation on social media is challenging. One intervention that has recently gained great attention is X/Twitter’s Community Notes (previously known as “Birdwatch”). Community Notes is a crowdsourced fact-checking approach that allows users to write textual notes to inform others about potentially misleading posts on X/Twitter. Yet, empirical evidence regarding its effectiveness in reducing engagement with misinformation on social media is missing. In this paper, we perform a large-scale empirical study to analyze whether the introduction of the Community Notes feature and its roll-out to users in the U.S. and around the world have reduced engagement with misinformation on X/Twitter in terms of retweet volume and likes. We employ Difference-in-Differences (DiD) models and Regression Discontinuity Design (RDD) to analyze a comprehensive dataset consisting of all fact-checking notes and corresponding source tweets since the launch of Community Notes in early 2021. Although we observe a significant increase in the volume of fact-checks carried out via Community Notes, particularly for tweets from verified users with many followers, we find no evidence that the introduction of Community Notes significantly reduced engagement with misleading tweets on X/Twitter. Rather, our findings suggest that Community Notes might be too slow to effectively reduce engagement with misinformation in the early (and most viral) stage of diffusion. Our work emphasizes the importance of evaluating fact-checking interventions in the field and offers important implications to enhance crowdsourced fact-checking strategies on social media.

- 4.1 Introduction 59
- 4.2 Background and Related Work 62
 - 4.2.1 Misinformation Interventions 62
 - 4.2.2 Community Notes on X/Twitter 63
- 4.3 Materials and Methods . 64
 - 4.3.1 Data Collection 64
 - 4.3.2 Variable Extraction 65
 - 4.3.3 Empirical Models 67
- 4.4 Empirical Analysis 69
 - 4.4.1 Fact-Checking via Community Notes 69
 - 4.4.2 Analysis at the Feature Roll-Out Level (DiD) . . 71
 - 4.4.3 Placebo Analysis 74
 - 4.4.4 Analysis at the Note Display Level (RDD) . . . 78
 - 4.4.5 Confounding Analyses . 80
- 4.5 Discussion 84
 - 4.5.1 Research Implications . . 85
 - 4.5.2 Practical Implications . . 87
 - 4.5.3 Limitations and Future Research 87
- 4.6 Conclusion 88

Publication

Y. Chuai, H. Tian, N. Pröllochs, and G. Lenzini, ‘Did the roll-out of Community Notes reduce engagement with misinformation on X/Twitter?’ *Proceedings of the ACM on Human-Computer Interaction*, vol. 8, no. CSCW2, pp. 1–52, 2024.

4.1 Introduction

Misinformation refers to false or misleading information that is contradicted by empirical evidence or is inconsistent with common or expert consensus [1, 7, 8]. The spread of misinformation has become a pervasive problem in the digital age, particularly on popular social media platforms such as X (formerly known as Twitter), TikTok, and Facebook. Online engagement with misinformation poses significant threats to the stability of society and contributes to contentious events across various domains, including, but not limited to, politics, finance, and public health [1, 6, 10]. During political elections, misinformation can influence voter opinions and behaviors, potentially undermining the democratic process [16, 18, 21]. In financial markets, misinformation can lead to market volatility and financial losses for investors and businesses. For example, 130\$ billion in stock value was wiped out in minutes following a false tweet indicating

that an “explosion” injured Barack Obama in 2013 [22]. Online misinformation can even pose severe risks to public health. For instance, during the COVID-19 pandemic, misinformation contributed to individuals’ hesitancy and reluctance to accept vaccinations, resulting in a heightened risk of infection [3, 109]. Consequently, mitigating user engagement with online misinformation is crucial and urgent.

To reduce engagement with online misinformation, various interventions have been proposed [1, 43]. These interventions can be broadly divided into two categories, namely, prebunking and debunking [1, 7, 43, 166]. Prebunking indicates pre-emptive interventions that aim to help individuals identify and resist misinformation they may subsequently encounter [1]. For example, a popular strategy is redirection, which focuses on directing users’ attention towards trustworthy information sources [1, 166]. While prebunking interventions can assist individuals in developing resistance to misinformation in a relatively broad context, they remain inadequate to counteract specific instances of misinformation that can quickly become viral and cause concern [1]. The other category of interventions, debunking, aims to fact-check the specific misinformation after its creation. Debunking is frequently carried out by third-party organizations, such as Snopes and PolitiFact, which recruit professional fact-checkers and subject experts to ensure the reliability of the verdicts. However, due to the limited number of human experts, it’s difficult to implement professional fact-checking in a timely and scaled manner. As an alternative, automatic fact-checking based on machine learning models can process large numbers of posts at a high speed. However, machine learning typically shows low performance in fact-checking real-world social media content due to the lack of counter-evidence in training data [58].

In order to address the drawbacks of the existing approaches to debunking, recent research has proposed to employ non-expert fact-checkers in the crowd to fact-check social media posts at scale [46, 54]. Though individual assessments can have bias and noise, the aggregated judgment of even relatively small crowds has been found to be comparable to those from expert fact-checkers [51, 52, 96]. Informed by these research findings, X/Twitter has launched its Community Notes platform (previously known as “Birdwatch”), which is the first large-scale effort of crowd-sourced fact-checking deployed in practice [54, 67]. The Community Notes platform allows users to identify tweets they believe are misleading or not misleading and write textual notes that provide context to the tweet. The Community Notes platform also features a rating system that allows users to rate the helpfulness of the fact-checks. If a note receives a high helpfulness score, it is displayed directly on the fact-checked tweets, visible to all users on X/Twitter. In an attempt to foster transparency, X/Twitter makes all Community Notes contributions available for analysis by researchers [167]. This offers a unique opportunity to analyze how users interact with this misinformation intervention in a real-world setting.

While the Community Notes feature has garnered considerable attention lately, an understanding of whether it is effective in reducing engagement (*e.g.*, retweet volume, likes) with misinformation on X/Twitter is missing. Previous research has shown that the fact-checks carried out on Community Notes are relatively accurate and trustworthy [54, 56, 96], and that Community Notes contributors focus on content posted by individuals with whom they hold opposing political views [50]. Moreover, during the pilot phase of the Community Notes feature, X/Twitter conducted survey experiments to show that note annotations were, to a large extent,

perceived as helpful by other users [67]. Users exposed to notes were also less likely to retweet or like misleading posts [67]. Additionally, recent research has analyzed the retweet volume (*i.e.*, the engagement in sharing) of posts that were fact-checked via Community Notes [96]. Contrary to misinformation fact-checked by third-party fact-checking organizations [14], community fact-checked misinformation received fewer retweets than truthful posts during the pilot phase of Community Notes, which can be partially attributed to the fact that community contributors were more likely to fact-check high-influence accounts with a large number of followers [96]. However, the key question of whether Community Notes is an effective intervention to reduce engagement with misinformation on social media has, so far, not been studied. This is the objective of this work.

In this paper, we conduct a comprehensive empirical study to investigate the effect of the Community Notes feature on user engagement with misinformation on X/Twitter in terms of retweet volume and likes. Our study covers all notes and associated tweets posted on X/Twitter between July 1, 2021, and April 6, 2023. This observation period includes two important intervention dates: October 6, 2022, and December 11, 2022. The first intervention occurred on October 6, 2022, when Community Note’s pilot phase ended and helpful notes became visible to users in the U.S. [168]. The second intervention occurred on December 11, 2022, when X/Twitter expanded Community Notes program to the rest of the world. This longitudinal dataset provides us with a unique opportunity to examine the impact of the roll-out of Community Notes on the engagement with misinformation on X/Twitter. In our analysis, we first examine the frequency of notes per tweet and the time it takes for notes to be created and displayed to users. Second, we use the tweets posted before and after the interventions to longitudinally examine the treatment effect with Difference-in-Differences (DiD) models at the feature roll-out level. This approach enables us to gauge the relative reduction in engagement with misinformation after the roll-out of the Community Notes feature compared to the pre-roll-out period. Third, we conduct a placebo analysis to validate the results obtained from the DiD models. Fourth, we limit our analysis to the tweets posted after the interventions and use Regression Discontinuity Design (RDD) as a complementary method to test the discontinuity (*i.e.*, treatment effect) at the note display level. This approach allows us to measure the relative reduction in engagement with misinformation on which notes are displayed compared to misinformation on which notes are not displayed. Finally, we conduct a comprehensive robustness analysis and consider potential confounding factors from multiple levels to further consolidate the findings of this study.

Our study finds no evidence to suggest that the roll-out of the Community Notes feature to users in the U.S. and around the world reduced engagement with misinformation on X/Twitter in terms of retweet count and like count. This finding is confirmed by both DiD and RDD models. Although we observe a significant increase in the volume and speed of fact-checking, particularly for tweets from verified users with many followers, our results suggest that the response time of Community Notes, *i.e.*, from the time of tweet creation to the time of displaying notes, is still insufficient to reduce engagement with misinformation in the early (and most viral) stage of diffusion. Additionally, we observe a decline in engagement with fact-checked tweets – irrespective of the veracity – after the expansion of the Community Notes feature to a global audience. Interestingly, this decline disappears when we exclusively analyze the

tweets from the set of authors that have been fact-checked both before and after the expansion. This suggests that the decrease in engagement with community fact-checked tweets can be partially attributed to differences in the fact-checking targets. Altogether, our results highlight the importance of evaluating misinformation interventions within real social media environments. We provide an in-depth discussion on the theoretical and practical implications of our work, which furnishes valuable insights into improving crowdsourced fact-checking strategies on social media.

4.2 Background and Related Work

4.2.1 Misinformation Interventions

As concerns regarding the impact of online misinformation on online users continue to grow, many intervention strategies have been proposed to reduce engagement with misinformation and ensure that people remain accurately informed. These interventions can be broadly distinguished into two categories, namely, prebunking and debunking [1].

Prebunking is a kind of pre-emptive intervention and seeks to help people recognize and resist subsequently encountered misinformation [1, 7]. Prebunking interventions include factually correct information presentation, pre-emptive correction, digital media literacy training, and generic misinformation warning before the misinformation. Redirection, which focuses on shifting users' attention to trustworthy information sources, is a kind of prebunking intervention implemented in practice by social media platforms [169]. The prebunking approach provides a promising tool to act pre-emptively and help people build immunity to misinformation in a relatively general manner [1, 170, 171]. While prebunking interventions can assist individuals in developing resistance to misinformation in a relatively broad context, they remain inadequate to counteract specific instances of misinformation that can quickly become viral and cause concern [1].

Debunking, also known as fact-checking, involves responding to specific instances of misinformation after they have been exposed to demonstrate why they are false [1, 7, 43]. Reliable fact-checking plays a crucial role in reducing engagement with false or misleading information and promoting accurate information that can inform public opinion and decision-making. Fact-checking strategies can be implemented in various ways, including third-party organizations and automated models [43]. Of note, manual professional fact-checking carried out by experts is a time-consuming process. Verdicts are usually reached after a message has gone viral, and interventions can have only limited effect [172]. Moreover, this strategy is difficult to scale due to the limited number of professional fact-checkers. To realize large-scale fact-checking and reduce response time, automatic fake news detection becomes a promising direction. Many machine learning models are trained based on the features of content (*e.g.*, writing style) and context (*e.g.*, network structure) to detect misinformation automatically. Nevertheless, because training data often lacks counter-evidence, fully automated fact-checking remains unrealistic [58].

A novel approach that has gained attention recently is to leverage the collective intelligence of the crowd, *i.e.*, wisdom of crowd, by allowing individuals who receive online information to do the fact-checking at scale [46, 51, 54, 156, 172, 173]. This approach is based on the expectation

that non-expert fact-checkers can achieve comparable accuracy to that of experts through collective assessments [96]. While individual assessments of non-expert fact-checkers can be unreliable and inconsistent, harnessing the collective judgment of a crowd can be an effective method for identifying misinformation. Recent studies demonstrate that even small crowds can produce accuracy levels comparable to those of expert fact-checkers [5, 51, 52].

In practice, various countermeasures against misinformation have been implemented by news media and social media platforms. Based on the redirection strategy, TikTok implemented a feature where a banner with links to trustworthy information sources was attached to COVID-19 related videos [174]. Google Search and YouTube have implemented a knowledge panel feature, which provides users with high-level credible information related to their search terms [175]. Instagram and Facebook apply labels to posts based on third-party fact-checker ratings to provide users with more accurate information [176]. For example, Snopes and PolitiFact are two pioneering U.S.-based third-party fact-checking organizations. They typically employ trained journalists or subject matter experts to investigate and verify the accuracy of claims made in news articles, social media posts, and other forms of online content [156]. Also, the International Fact-Checking Network (IFCN) at Poynter Institute was launched to bring together fact-checkers from more than 100 organizations around the world and support the global fight against misinformation [20]. So far, however, such large-scale fact-checking efforts (and research) are mainly focused on the Global North. In the Global South, fact-checking efforts are still in its infancy and face additional challenges (*e.g.*, due to under-resourced languages, lack of fact-checking sources) [20, 177]. Ultimately, X/Twitter recently implemented the Community Notes feature as a significant step to combat misinformation on their platform [178]. This innovative feature marks the first large-scale effort of crowdsourced fact-checking deployed in practice [54].

Altogether, various misinformation interventions have been proposed by researchers or implemented by social media platforms [1, 54, 170, 174–176]. Yet, their effectiveness in reducing engagement with misinformation is oftentimes unclear. It is thus essential to evaluate the effectiveness of misinformation interventions directly on social media platforms.

4.2.2 Community Notes on X/Twitter

X/Twitter launched the pilot program of Community Notes on January 28, 2021, and limited its visibility to a small number of pilot participants. Thus, during the pilot phase, the community fact-checks were not expected to directly influence engagement with the fact-checked tweets [96]. The Community Notes feature was later expanded to all users in the U.S. on October 6, 2022, and to global users on December 11, 2022 [168].

In an attempt to foster transparency, X/Twitter makes all Community Notes contributions available for analysis by researchers [167]. Prior works have used the data to study the mechanisms and behaviors of raters on the platform [50, 167]. Note contributors write more notes for misleading than not misleading tweets [167], tend to focus on content posted by individuals with whom they hold opposing political views [50], and are more likely to fact-check posts from influential user accounts with many followers [96]. Furthermore, research has shown that fact-checks for tweets from influential accounts yield a lower level of consensus among the contributors [167] and that the assessments made

on Community Notes are comparable to those obtained from human experts [54]. Recent research has also used the fact-checked tweets on Community Notes to investigate the diffusion patterns of misinformation on X/Twitter [96, 179]. Contrary to misinformation fact-checked by third-party fact-checking organizations [14], community fact-checked misinformation received fewer retweets than truthful posts during the pilot phase of Community Notes [96]. Additionally, during the pilot phase, a test conducted by X/Twitter demonstrated that Community Notes could reduce users' intentions to share misinformation [67].

In sum, while the Community Notes feature has garnered considerable attention lately, it still remains unclear whether it is an effective tool in reducing engagement with misinformation among actual users on the social media platform. The half-life of a tweet is short: about 95% of tweets have no relevant impressions after two days, and it only takes about 79.5 minutes before half of impressions are created for a tweet [180]. Hence, even if Community Notes reduces sharing intentions, the time delay until a community note becomes visible may be too long to effectively reduce engagement with misinformation in the early (and most viral) stage of the diffusion. Further research is needed to determine the actual impact of Community Notes on the overall engagement with misinformation on X/Twitter.

4.3 Materials and Methods

To evaluate the effectiveness of Community Notes in reducing engagement with misinformation on X/Twitter, we first download all note contributions on the Community Notes platform. Second, through X/Twitter's Academic Research API, we collect the fact-checked tweets using the tweet IDs referenced in Community Notes. Third, we measure engagement using retweet and like counts, and extract a wide variety of additional variables that may affect the engagement with content on X/Twitter (e.g., the number of followers, emotions, topics) from both notes and tweets. Subsequently, we employ Difference-in-Differences (DiD) as the primary method to analyze whether the introduction of the Community Notes feature and its roll-out to users in the U.S. and around the world has reduced engagement with misinformation on X/Twitter (*i.e.*, at the feature roll-out level). Additionally, we utilize Regression Discontinuity Design (RDD) as a complementary method to examine the effectiveness of Community Notes at the note display level. The latter approach allows us to measure the relative reduction in engagement with misleading posts on which notes are displayed compared to those on which notes are not displayed.

4.3.1 Data Collection

All note contributions made on Community Notes are available for the public via a dedicated website.¹ The data used in our study was downloaded on April 8, 2023, and it covers all contributions up until April 6, 2023, such as notes, ratings, note status history, and user enrollment status. Specifically, the data includes 68,235 notes, 75,508 note status histories, 2,293,934 ratings, and 64,359 enrolled users. Notes deleted by users still retain their status histories and ratings. Each note contains its ID, referenced tweet ID, creation time, and classification (*i.e.*, "misinformed or potentially misleading" or "not misleading"). Each note status item

1: <https://twitter.com/i/communitynotes/download-data>

records the change of referenced note statuses, including “Needs More Ratings (NMR)”, “Currently Rated Helpful (CRH)”, and “Currently Rated Not Helpful (CRNH)”, which are determined by their helpfulness scores. All fact-checking notes on Community Notes start with the “Needs More Ratings” status until they receive at least 5 total ratings. To prevent one-side ratings and identify notes that are helpful to a wide range of people, Community Notes uses matrix factorization on the note-rater matrix to calculate note helpfulness scores. Notes marking tweets as misinformation (misinformed or potentially misleading) with a note helpfulness score of 0.40 and above earn the status of Helpful [67].

In addition to the contribution data, the note ranking algorithm code that X/Twitter runs in production is also released on GitHub.² We download the open-source code to calculate the helpfulness score for each note. As of April 13, 2023, the source code is in version 1.0 and uses a multi-model note ranking strategy. Three note ranking models, namely, Core, Expansion, and Coverage, run the same ranking algorithm but differ in their ranking areas or thresholds. The Core model is always the most authoritative one. After running the code, we get the note score for each note. For the non-deleted notes, 63,870 out of 68,235 (93.6%) are determined by the Core model.

2: <https://github.com/twitter/communitynotes/tree/main/sourcecode>

After collecting the Community Notes data, we identify the tweets referenced by the notes based on the associated 45,664 unique tweet IDs. Utilizing the X/Twitter Academic Research API, we map these tweet IDs to the corresponding source tweets and successfully retrieve 39,268 tweets, which corresponds to 86% of all tweets. Thus, the vast majority of the source tweets fact-checked via Community Notes are captured in our dataset. To address the concern that some potentially harmful misleading tweets with CRH notes may be removed by X/Twitter and cannot be fetched via IDs through the API [181], we specifically examine the subset of tweets with CRH notes. Out of 4,686 tweets with CRH notes, 3,872 (82.6%) are successfully retrieved. This retrieval rate for tweets with CRH notes is similar to the overall rate for all tweets, which indicates a consistent level of data retrieval across categories. Furthermore, it is important to emphasize that X/Twitter explicitly states that notes that are rated helpful appear directly on the source tweets and have no effect on the display of source tweets or enforcement of X/Twitter’s rules.³

3: <https://communitynotes.twitter.com/guide/en/about/faq>

We perform two filtering steps for our analysis. First, this study focuses on tweets posted in English and we filter out all source tweets that were not posted in the English language. This results in a total of 36,077 tweets authored by 15,348 unique users. Second, since Community Notes is a new feature introduced in January 2021 and requires time to gain contributors, we restrict our analysis to tweets posted after July 1, 2021. This allows for a period of six months for Community Notes to gain attention and for enough high-quality notes to be written and rated by the community. After filtering, we obtain 30,267 unique tweets posted between July 2021 and April 2023, which account for 83.9% of the original set of English tweets. These tweets serve as the basis for our analysis and are subsequently divided into multiple groups to study the effect of Community Notes on reducing engagement with misinformation on X/Twitter.

4.3.2 Variable Extraction

In this paper, we examine the effect of the roll-out of Community Notes on the engagement with misinformation on X/Twitter using two commonly-

used engagement metrics, namely (1) the retweet count and (2) the like count. Retweet count (*RetweetCount*) is a count variable denoting the number of retweets (not including quote tweets) a single tweet receives on X/Twitter. Tweets can reach a large number of people through retweets. Hence retweet count is the most used metric to measure the virality of posts on X/Twitter [10, 14, 96]. Additionally, like count (*LikeCount*) is a count variable that measures how many users on X/Twitter have indicated that they appreciate or agree with the content of a tweet [67]. The variables of *RetweetCount* and *LikeCount* are directly collected from the fields of public metrics embedded in the tweet objects.

Moreover, we use the X/Twitter Academic Research API to collect a wide variety of variables that may affect engagement with tweets on X/Twitter:

- ▶ **WordNum:** A count variable indicating the number of words in a single tweet. Text length can have a positive effect on the perceived text quality [10, 182]. To calculate the text length, we remove all links, stopwords, and non-alphabetic characters from the tweet text. We also remove the “#” prefix from hashtags and retain the text. Additionally, we remove the “@” prefix from user accounts mentioned in the content, treating the usernames the same as the names of individuals mentioned in the content. *WordNum* then corresponds to the number of remaining words in each tweet.
- ▶ **Media:** A dummy variable indicating whether the tweet contains media elements, such as pictures or videos (= 1; otherwise 0). Media elements can enhance user engagement with the corresponding information [127].
- ▶ **Hashtag:** A dummy variable indicating whether the tweet contains one or more hashtags (= 1; otherwise 0). By using one or more hashtags, users can link their tweets to larger conversations or hot topics. This can increase the visibility of their tweets [82].
- ▶ **Mention:** A dummy variable indicating whether the tweet contains one or more mention symbols (= 1; otherwise 0). Users can mention other users in the content of tweets, and this can also affect engagement with the tweets [82].
- ▶ **AcctAge:** A count variable indicating the age of the account (in days) at the time of data collection.
- ▶ **Verified:** A dummy variable indicating whether the author is verified (= 1; otherwise 0).
- ▶ **Followers:** A count variable indicating the number of followers of the author.
- ▶ **Followees:** A count variable indicating the number of followees of the author.

In addition to the above variables, engagement may also be affected by emotions and topics in the fact-checked tweets [10, 14, 34, 81, 98, 183]:

- ▶ **Emotions:** To control the possible effects of emotions on the engagement with information, we consider six fine-grained basic emotions, namely, anger, disgust, joy, sadness, fear, and surprise [10, 14]. We first perform standard preprocessing steps (see *WordNum* variable). Subsequently, we use Spacy, an Industrial-Strength Natural Language Processing library in Python, to lemmatize the words. Next, we count the frequency of words that belong to each of the emotions in each tweet based on the commonly used NRC emotion lexicon [163]. As a result, 81.2% of all tweets contain at least one emotion word. We then calculate the weight of each emotion in each tweet according to the number of corresponding emotion words. Notably, tweets without

any emotion words are still included in the analysis, with the weights of all emotions being 0.

- **Topics:** X/Twitter has already labeled the corresponding domains that each tweet belongs to in the collected dataset. We extract all the domains in the fact-checked tweets and classify them into six general topics, *i.e.*, Entertainment, Business & Finance, Politics, Science & Technology, Emergency Events, and Ongoing News Story. The topic list is shown in Table S4.1. For each topic, we have a corresponding dummy variable to denote whether the tweet belongs to it (= 1; otherwise 0). Each tweet can have multiple topics. Thus, the sum of the topic dummy variables for each tweet is not constrained to equal one.

Descriptive statistics for the extracted variables are reported in Table S4.2.

4.3.3 Empirical Models

We evaluate the effectiveness of Community Notes using two popular quasi-experimental methods. First, we use tweets posted before and after the interventions to longitudinally examine the treatment effect with *Difference-in-Differences (DiD)* models at the feature roll-out level. This approach enables us to gauge the additional reduction in engagement with misinformation on X/Twitter after the roll-out of Community Notes feature compared to the pre-roll-out period. Second, we limit our analysis to tweets posted after the interventions, and use *Regression Discontinuity Design (RDD)* as a complementary method to test the discontinuity (*i.e.*, treatment effect) at the note display level. Here, the treatment effect indicates the extra reduction in engagement with misinformation on which notes are displayed compared to the engagement with misinformation on which notes are not displayed.

Difference-in-Differences (DiD) model. We use Difference-in-Differences (DiD, ref. [184]) to estimate the treatment effect of Community Notes in reducing engagement with misinformation on X at the feature roll-out level. Longitudinally, the DiD model assumes that the treatment group (the group affected by the treatment) and control group (the group not affected by the treatment) follow parallel trends in the absence of the intervention.⁴ This assumption allows for the estimation of the average treatment effect on the treated group (ATT). Specifically, the model compares the changes in an outcome variable over time for a treatment group to the changes in the same outcome variable over time for a control group. Importantly, the term “intervention” in this study refers to the “roll-out” of the Community Notes feature, and the term “treatment” refers to the display of notes on the corresponding source tweets.

There are two dummy variables associated to the group level and time period level in the DiD model: (i) The group-specific dummy variable *Treated* indicates the treatment group (= 1) or control group (= 0). (ii) The time-specific dummy variable *Vis* indicates the pre-intervention (= 0) or post-intervention period (= 1). Comparing the differences in engagement between the two groups before and after the intervention, the DiD model can help identify the treatment effect of the intervention by estimating the interaction term $Treated \times Vis$. Note that the treatment group includes tweets that receive one or more CRH notes, regardless of the delay in note display (*Delay*). The control group includes the tweets that did not receive any CRH notes. Then, we can use DiD to estimate the extra reduction of average engagement in the treatment group after

4: To ensure a parallel trend before the intervention between treatment group and control group, we later employ propensity score matching (see details in Section 4.4.2). Additionally, given that tweets treated after the intervention are not the same as those treated before the intervention, we consider confounding factors (*i.e.*, control variables, *e.g.*, content and author characteristics) that may affect the outcome alongside the roll-out of Community Notes feature (see details in Section 4.4.5).

the intervention, compared to the baseline before the intervention and relative to the control group.

Formally, our DiD model is specified as:

$$\log(E(\text{Outcome}_{it}|x_{it}^*)) = \beta_0 + \beta_1 \text{Treated}_i + \beta_2 \text{Vis}_t + \beta_3 \text{Treated}_i \times \text{Vis}_t + \alpha' x_{it}, \quad (4.1)$$

where Outcome_{it} indicates the retweet count (RetweetCount_{it}) or like count (LikeCount_{it}) of tweet i posted at time t . Because retweet count and like count are all non-negative count variables with overdispersion, we employ a negative binomial regression. Treated_i is the group-specific dummy variable indicating whether the tweet i belongs to the treatment group ($= 1$) or control group ($= 0$). Vis_t is the time-specific dummy variable indicating whether the tweet i is posted after the roll-out of Community Notes ($= 1$) or not ($= 0$). The interaction term, $\text{Treated}_i \times \text{Vis}_t$, is difference-in-differences estimation and denotes the treatment effect (*i.e.*, the effect attributed to the roll-out of Community Notes). x_{it} is a vector including control variables constructed in Section 4.3.2. x_{it}^* indicates all the independent variables. β_0 is the intercept. For the sake of interpretability, we z-standardize all continuous variables in x_{it}^* .

Regression Discontinuity Design (RDD) model. Given that the detailed engagement timelines before and after the display of CRH notes are not available in our dataset, we cannot use DiD to estimate the treatment effect of Community Notes program at the note display level (*i.e.*, the treatment effect attributed to the display of notes). However, we can use Regression Discontinuity Design (RDD, ref. [184]) as a complementary method to conduct an analysis of the effectiveness of Community Notes at the note display level in terms of aggregated engagement. RDD is a quasi-experimental research design that aims to estimate the effect of a treatment by exploiting a discontinuity in the relationship between a running variable and an outcome variable at a specific threshold or cutoff point. The key rationale behind this is that the expected outcomes from samples with close values on a running variable should exhibit smooth continuity around the cutoff point without the treatment. Therefore, in an RDD model, samples are assigned to the treatment or control group based on their values on a running variable, and the treatment effect is estimated by comparing the outcomes of samples just above and just below the cutoff point [185–187]. In our analysis, NoteScore is the running variable that determines whether the notes are displayed (*i.e.*, CRH notes or non-CRH notes). We use the values of NoteScore to select tweets just above (treatment group, $\text{Treated} = 1$) and just below (control group, $\text{Treated} = 0$) the threshold in a narrow window. By estimating the group-specific dummy variable Treated , the RDD model can identify the discontinuity jump (treatment effect) from control group to treatment group.

Formally, our RDD model is specified as:

$$\log(E(\text{Outcome}_i|x_i^*)) = \beta_0 + \beta_1 \text{Treated}_i + \beta_2 \text{Delay}_i + \beta_3 \text{NoteScore}_i + \alpha' x_i, \quad (4.2)$$

where Outcome_i indicates retweet count or like count of tweet i . Different from the DiD model, the RDD model does not take into account the longitudinal time periods before and after the intervention, but rather focuses on the running variable NoteScore_i that determines the display of the note. The cutoff point for NoteScore_i is set at 0.40, indicating that only notes with a helpfulness score of 0.40 or above are eligible to be displayed on the corresponding tweets. Note that in the RDD model,

we only consider tweets with notes decided by the Core model (v1.0) to keep the same standard on the running variable. The selection of the window around the cut-off point is specified in the analysis (see Section 4.4.4). $Treated_i$ captures the discontinuity (*i.e.*, treatment effect) at the point of cutoff. $Delay_i$ is the time duration (days) between the posting time of tweet i and the point at which its associated note is displayed to the users. x_i is a vector including control variables constructed in Section 4.3.2. x_i^* indicates all the independent variables. β_0 is the intercept. We again z -standardize all continuous variables in x_i when fitting the RDD model.

Note that in both models, *i.e.*, DiD and RDD, the treatments are all indicated by the note helpfulness scores (*i.e.*, *NoteScore*) that decide upon the note statuses. However, the two methods incorporate *NoteScore* and *Delay* slightly differently. Specifically, at the roll-out level, the DiD model does not incorporate the continuous variable *NoteScore* as it instead uses a binary treatment indicator (*i.e.*, *Treated*). This binary indicator categorizes tweets into treated and control groups based on the presence or absence of helpful notes, providing a clear distinction for the evaluation of engagement levels before and after the intervention. In contrast, at the note display level, RDD requires specific values of *NoteScore* to define the cutoff point, set the selection window, and construct treatment and control groups. A similar rationale applies to the variable *Delay*. At the roll-out level, DiD generally evaluates the average engagement level before and after the intervention, comparing the treatment and control groups. The roll-out time of Community Notes feature is thus constant across each treated tweet, irrespective of the delay in note display. The RDD model, on the other hand, focuses on the note display level. The timing of the treatment thus varies among the treated tweets. Consequently, *Delay* needs to be explicitly included in the RDD model.

4.4 Empirical Analysis

We begin our empirical analysis by summarizing the fundamental changes in notes and fact-checked tweets upon the expansion of Community Notes to the U.S. and the rest of the world (Section 4.4.1). We then create treatment and control groups covering time periods before and after the expansions, followed by applying the DiD model to study the effectiveness of Community Notes at the feature roll-out level (Section 4.4.2). Subsequently, we present a placebo analysis with different combinations of treatment and control groups (Section 4.4.3). Afterwards, we construct treatment and control groups based on the running variable of note helpfulness, and employ the RDD model to estimate the treatment effect at the note display level (Section 4.4.4). Finally, we consider potential confounding factors across multiple levels to further consolidate the findings of this study (Section 4.4.5).

4.4.1 Fact-Checking via Community Notes

With $\times$ expanding Community Notes to the U.S. and the world, we first perform an initial analysis on the development of the Community Notes feature in terms of contributors, fact-checking notes, and tweets. Of the 64,359 contributors in total, 48,927 contributors (76%) joined after the global expansion of Community Notes, marking a more than 24-fold increase in the number of contributors during the “Visible World” phase

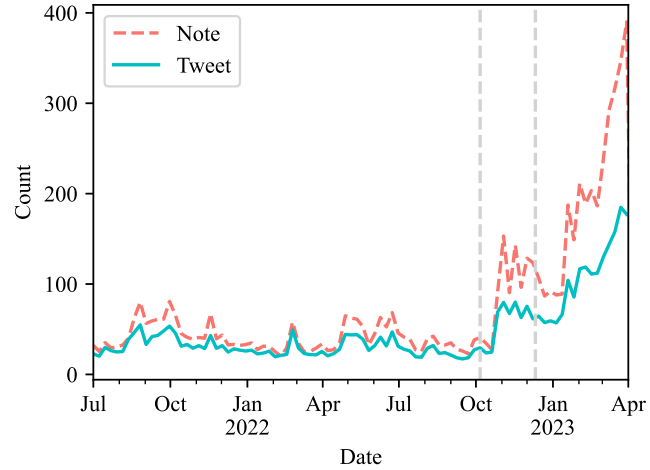


Figure 4.1: Daily counts of created notes and fact-checked tweets, smoothed by a 7-day moving average. There are two grey vertical dash lines. The first (left) is on October 6, 2022 (Community Notes feature visible in the U.S.), and the second (right) is on December 11, 2022 (Community Notes feature visible to the world).

(visible to the world) compared to the “Visible US” phase (visible only in the U.S.). Simultaneously, we observe a substantial surge in the daily counts of notes and fact-checked tweets, especially after the expansion to the world (Fig. 4.1). In addition, we find that the number of notes grows faster than the number of fact-checked tweets. The ratio of notes to tweets increased from 1.37 during the Invisible phase to 1.62 in the “Visible US” phase, and further to 1.79 in the “Visible World” phase. On average, each tweet received 1.58 notes, with 32.2% of tweets being fact-checked by more than one note.

We now consider the relationships between the volume of notes and various tweet characteristics (see variables in Section 4.3.2). As the majority of tweets receive only a single note, we employ a logistic regression model to analyze the characteristics of tweets that are more likely to receive multiple notes (coded as 1, otherwise 0). Moreover, we model the rating number (*i.e.*, the number of ratings a note receives) with a negative binomial regression. We find that tweets from verified high-influence users with many followers tend to get more notes and ratings (see the coefficients of *Verified* and *Followers* in Columns (1) and (2) of Table 4.1). Tweets with media elements, such as pictures and videos, also have a greater probability of receiving notes and ratings (see the coefficients of *Media* in Columns (1) and (2) of Table 4.1).

Additionally, we analyze the response time (delay) that notes take to be created and displayed to the users. As shown in Fig. 4.2(a), the delay from tweets to notes creation has decreased significantly from 5.21 days (before the roll-outs) to 3.08 days (Visible US, $t = 3.511, p < 0.001$), and to 1.82 days (Visible World, $t = 7.763, p < 0.001$), on average. When moving the focus to the response time that notes take to be displayed, we find that the delay also decreased (Fig. 4.2(b)); however, the magnitude of the decrease is relatively small ranging from 2.85 days (Visible US) to 2.23 days (Visible World, $t = 2.286, p < 0.05$). Notably, these response times are significantly longer than the typical half-life of tweets.⁵ This indicates that the Community Notes program’s response time may not be fast enough to display notes during the early stages of diffusion, potentially limiting its effectiveness in significantly reducing engagement with misinformation on X.

We also model the response time with linear regression models and find that tweets from verified accounts tend to receive notes faster and with a shorter delay between the creation date and note display (see the coefficients for *Verified* in Columns (3) and (4) of Table 4.1).

5: Previous research found that the median half-life of a tweet, *i.e.*, the time it takes before half of all impressions are created, is about 79.5 minutes [180].

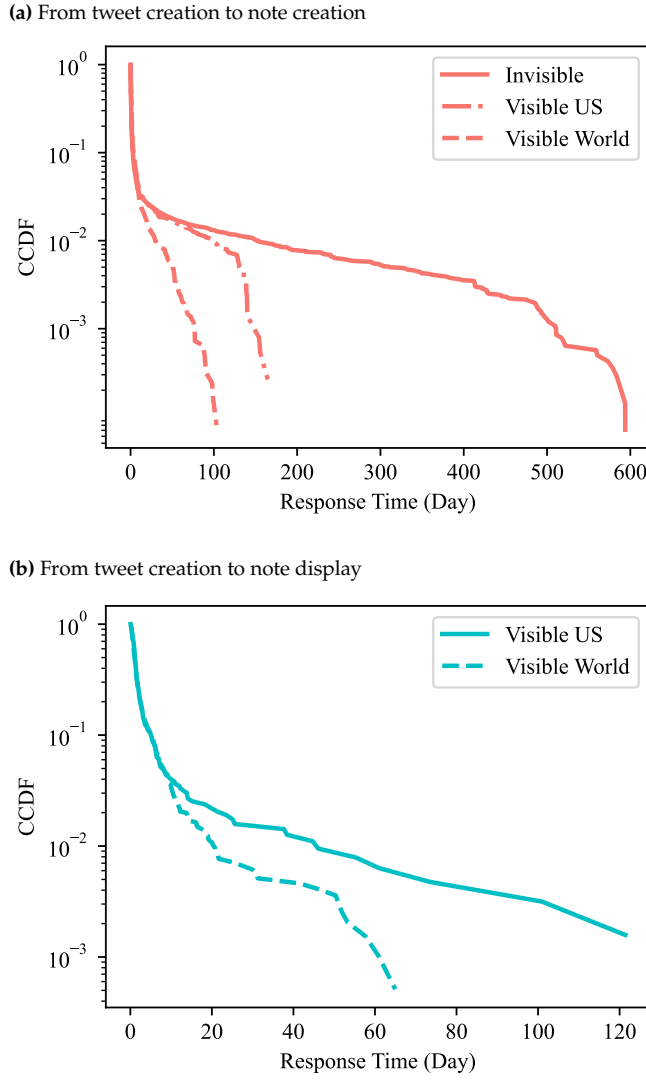


Figure 4.2: Complementary cumulative distribution functions (CCDFs) of the delays. (a) Delay (response time) between the time of tweet creation and the time of note creation. (b) Delay (response time) between the time of tweet creation and the time when the note is displayed to users.

In sum, our initial analysis reveals a rapid adoption of Community Notes, marked by a significant increase in contributors, notes, and fact-checked tweets, especially after the global expansion. Verified users with many followers and tweets with media elements tend to receive more notes and ratings. Despite a decrease after the roll-outs, the response time from tweet creation to note display is still significantly longer than the typical half-life of tweets, potentially limiting the effectiveness of Community Notes in reducing overall engagement with misinformation on $\mathbb{X}$.

4.4.2 Analysis at the Feature Roll-Out Level (DiD)

In this section, we employ Difference-in-Differences (DiD) to analyze whether the introduction of the Community Notes feature and its roll-out to users in the U.S. and around the world has reduced engagement with misinformation on $\mathbb{X}$ (*i.e.*, at the feature roll-out level). For this purpose, we categorize tweets into three groups based on their note statuses and classifications. The first group, F-CRH, consists of tweets with notes that indicate they are potentially misleading and are currently rated helpful. Notes with CRH status are in the highest priority and displayed on the corresponding tweets directly. Misleading tweets (*i.e.*,

Table 4.1: Regression results for response number and time. Regression models with robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

	Response Number		Response Time	
	(1) Note	(2) Rating	(3) Creation	(4) Visible
<i>WordNum</i>	0.042** (0.015)	0.020 (0.020)	0.485** (0.179)	0.215 (0.138)
<i>Media</i>	0.055* (0.028)	0.155*** (0.034)	0.317 (0.329)	0.155 (0.293)
<i>Hashtag</i>	-0.268*** (0.048)	-0.444*** (0.052)	0.196 (0.570)	-0.031 (0.294)
<i>Mention</i>	-0.614*** (0.036)	-0.680*** (0.054)	0.323 (0.401)	-0.084 (0.336)
<i>AcctAge</i>	-0.029* (0.014)	0.026 (0.016)	-0.224 (0.137)	-0.157 (0.121)
<i>Verified</i>	0.781*** (0.032)	0.756*** (0.038)	-3.219*** (0.377)	-1.218*** (0.283)
<i>Followers</i>	0.261*** (0.016)	0.277*** (0.013)	0.084 (0.173)	0.047 (0.169)
<i>Followees</i>	0.025* (0.012)	-0.022** (0.008)	0.015 (0.059)	-0.012 (0.051)
Emotions	✓	✓	✓	✓
Topics	✓	✓	✓	✓
Periods	✓	✓	✓	✓
<i>Intercept</i>	-1.832*** (0.039)	2.353*** (0.042)	8.042*** (0.564)	3.749*** (0.472)
Observations	30,267	30,267	30,267	2,586

misinformation, false information) in the F-CRH group contain at least one CRH note and can also contain other NMR or CRNH notes. The number of tweets in the F-CRH group is 3,387. As of April 13, 2023, Community Notes has different status control rules for notes indicating misleading and not misleading. Notes marking corresponding tweets as “not misleading” have no CRH status, even if they have enough ratings and high helpfulness scores. Considering this subset of tweets, we have the second group, T-NMR, which includes tweets with NMR notes indicating that they are not misleading (*i.e.*, true information) and have at least 5 ratings. Excluding tweets in the F-CRH group, the T-NMR group has 5,602 tweets with a high probability of being true. The third group, F-NCRH, includes tweets with notes indicating that they are potentially misleading but do not have a CRH status. Excluding tweets in the F-CRH and T-NMR groups, the F-NCRH group contains 20,669 unique tweets.

Community Notes implemented two interventions to change note visibility settings during the program. On October 6, 2022, the first intervention made notes visible to all users in the U.S. [168]. Two months after the first intervention, Community Notes expanded note visibility to the whole world on December 11, 2022, marking the second intervention. As of the date of data collection, the second intervention had been in effect for almost four months. Specifically, the two intervention periods are *VisUS* (06-10-2022 to 10-12-2022) and *VisWorld* (11-12-2022 to 06-04-2023). To perform parallel tests and longitudinal comparisons between the treatment and control groups, we divide the time before the interventions into five three-month periods, namely, *Pre5* (01-07-2021 to 30-09-2021), *Pre4* (01-10-2021 to 30-12-2021), *Pre3* (01-01-2022 to 31-03-2022), *Pre2* (01-04-2022 to 30-06-2022), and *Pre1* (01-07-2022 to 05-10-2022). Notably, we treat the roll-out of the Community Notes to the U.S. and the rest of

the world as two interventions. The reason is that there were significant changes in contributors, fact-checking notes, and tweets along with the global availability of Community Notes (see Section 4.4.1). Furthermore, even though there was no distinction in visibility to U.S. users between the two interventions, the Community Notes feature was still evolving and gaining traction among U.S. users. The impact of Community Notes during different phases may thus exhibit a different effect – even for U.S. users. Based on these reasons, we consider the two interventions separately to explore whether the effect of Community Notes after the second intervention is larger than that after the first intervention and to give more detailed estimations.⁶

Longitudinally, we take F-CRH as the treatment group and F-NCRH as the control group. CRH notes on tweets within the F-CRH group are visible to the U.S. and the whole world respectively during the two intervention periods (as illustrated in Fig. S4.1(a)). First, we conduct propensity score matching (PSM) to match F-CRH and F-NCRH groups according to the pre-defined variables in Section 4.3.2. With the caliper of 0.001, “common” support, and “noreplacement” settings, we only discard 38 out of 24,056 tweets (0.15%) to achieve all biases lower than 0.05 across the variables derived from tweets and authors. Second, we test the parallel trends between F-CRH and F-NCRH groups. Before the interventions, the changes in retweet count and like count within F-CRH and F-NCRH groups present a parallel trend and have no statistically significant difference (Fig. S4.2(a)). Ultimately, our investigation proceeds to analyze the impact of the interventions on the dissemination of false or misleading information. Regarding the first intervention, where Community Notes was expanded to the U.S., we consider the pre-intervention period as the three-month period leading up to October 6, 2023 (*Pre1*); while the two-month period from October 6, 2023 to December 11, 2022 is taken as the post-intervention period (*VisUS*). The coefficients of $FCRH \times Vis$ are all significantly negative in Columns (1) and (3) of Table 4.2 (*coef.* = -0.439 , $p < 0.01$ in Column (1); *coef.* = -0.403 , $p < 0.05$ in Column (3)). This indicates that F-CRH tweets received 35.5% fewer retweets (95% CI: $[-53.1\%, -11.4\%]$) and 33.2% fewer likes (95% CI: $[-51.9\%, -7.2\%]$) compared to the baseline before the first intervention and relative to the F-NCRH tweets. Community Notes thus seems to reduce the engagement with misinformation after the CRH notes are visible to the U.S. from the perspectives of retweet count and like count.

Continuing our investigation, we further scrutinize the effect of the second intervention, *i.e.*, the global expansion of the Community Notes feature. Here, the pre-intervention phase is defined as the two-month period ranging from October 6, 2023 to December 11, 2022 (*VisUS*). The post-intervention phase, on the other hand, is characterized by the duration of four months starting on December 11, 2022 and ending on April 6, 2023 (*VisWorld*). As shown in Columns (2) and (4) of Table 4.2, in the second intervention, the coefficients of $FCRH \times Vis$ for both retweet count and like count are found to be statistically non-significant. This suggests that the impact of Community Notes feature on reducing engagement with false or misleading information after the second intervention does not exhibit a significant difference compared to that observed after the first intervention. Fig 4.3 displays the variations in the retweet and like counts within the F-CRH and F-NCRH groups across both the pre- and post-intervention periods. During the first intervention, there are slight decreases in both retweet count and like count within the F-CRH group. In the absence of the first intervention, it is anticipated that the engagement with F-CRH tweets should go like the lighter red dash line

6: As an additional check, we also repeated our analysis without distinguishing between the two expansions of the Community Notes feature (*i.e.*, with a single treatment framework). The results are consistent with our two-treatment examination framework. The additional results are reported in Table S4.3.

Table 4.2: DiD regression results for retweet count and like count within F-CRH (treatment) and F-NCRH (control) groups. Negative binomial regression models with robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

	Retweet Count		Like Count	
	(1) Pre-US	(2) US-World	(3) Pre-US	(4) US-World
<i>FCRH</i>	1.172*** (0.132)	0.855*** (0.120)	1.212*** (0.139)	0.944*** (0.115)
<i>Vis</i>	0.177* (0.080)	−0.761*** (0.065)	0.202* (0.085)	−0.809*** (0.072)
<i>FCRH × Vis</i>	−0.439** (0.162)	−0.123 (0.133)	−0.403* (0.168)	−0.127 (0.132)
<i>WordNum</i>	0.081* (0.038)	0.134*** (0.031)	−0.071 (0.042)	−0.020 (0.033)
<i>Media</i>	0.515*** (0.078)	0.687*** (0.057)	0.453*** (0.084)	0.509*** (0.063)
<i>Hashtag</i>	−0.872*** (0.150)	−0.722*** (0.096)	−1.070*** (0.141)	−0.970*** (0.096)
<i>Mention</i>	−1.188*** (0.102)	−1.291*** (0.072)	−1.163*** (0.110)	−1.302*** (0.075)
<i>AcctAge</i>	−0.021 (0.034)	−0.018 (0.025)	−0.002 (0.033)	0.012 (0.028)
<i>Verified</i>	0.596*** (0.083)	0.821*** (0.061)	0.502*** (0.083)	0.783*** (0.065)
<i>Followers</i>	0.159*** (0.015)	0.179*** (0.013)	0.199*** (0.014)	0.221*** (0.012)
<i>Followees</i>	0.037 (0.036)	0.028 (0.016)	0.041 (0.039)	0.030* (0.012)
Emotions	✓	✓	✓	✓
Topics	✓	✓	✓	✓
<i>Intercept</i>	7.062*** (0.099)	6.899*** (0.080)	8.820*** (0.109)	8.658*** (0.090)
Observations	5,019	11,821	5,019	11,821

(according to the coefficients of *Vis* in Columns (1) and (3) of Table 4.2). The differences between the two red dash lines in Figs. 4.3(a) and (b) during the first prevention period can be regarded as the treatment effects. In the second intervention, there is a significant decrease in the engagement with tweets throughout both F-CRH and F-NCRH groups (according to the coefficients of *Vis* in Columns (2) and (4) of Table 4.2). However, the non-significant coefficients of *FCRH × Vis* for both retweet count and like count indicate that no special treatment effect on the F-CRH group is captured (Columns (2) and (4) of Table 4.2).

In summary, our findings so far indicate that the first intervention, involving the expansion of Community Notes to the U.S., is linked to decreases in engagement with misinformation, as evidenced by reduced retweet and like counts. The second intervention, *i.e.*, the global roll-out, does not yield a significant increase in the treatment effect as compared to the first intervention.

4.4.3 Placebo Analysis

We execute two placebo tests to further explore the effect of Community Notes in reducing engagement with misinformation. The special group, T-NMR, is taken into the placebo tests as a mediating group. In the first test, we replace F-CRH with T-NMR as the virtual treatment group and compared it to F-NCRH as the control group. In the second test, we use

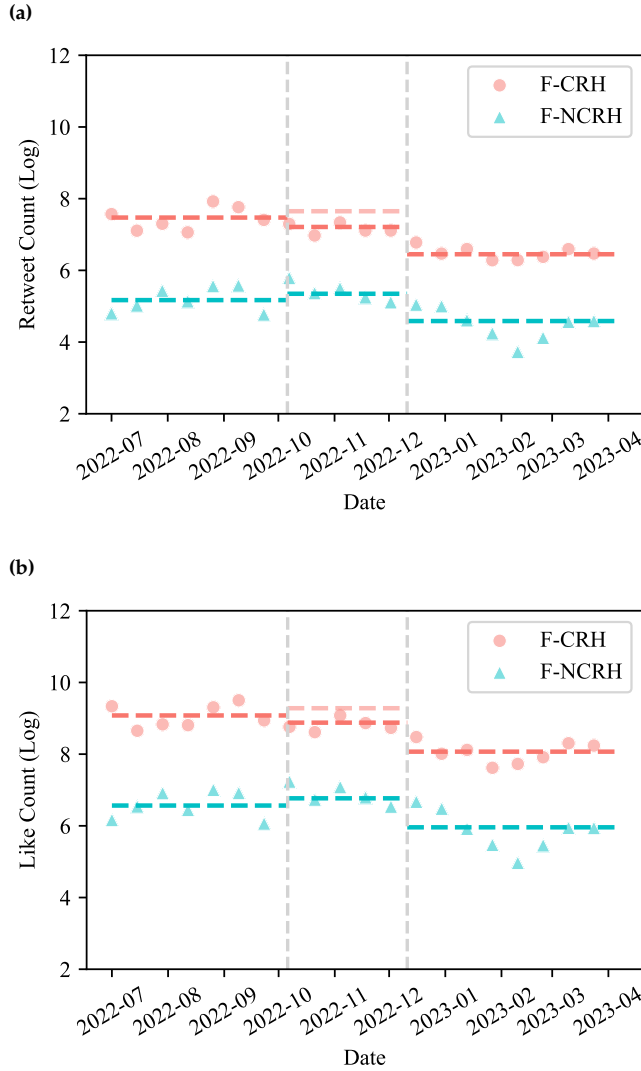


Figure 4.3: Two-week rolling averages of retweet count and like count in treatment (F-CRH) and control (F-NCRH) groups from 2022-07 to 2023-04. The two vertical grey dash lines in each figure indicate two dates, October 6, 2022 (left) and December 11, 2022 (right), when Community Notes was expanded to the U.S. and the world respectively. (a) The points are 2-week rolling averages of retweet count. The horizontal dash lines represent the averages of retweet count during different time periods, and illustrate the results of DiD models in Columns (1) and (2) of Table 4.2. The lighter red dash line during *VisUS* denotes the estimated average without the treatment effect ($FCRH \times Vis = -0.439, p < 0.01$) in Column (1) of Table 4.2. (b) The points are 2-week rolling averages of like count. The horizontal dash lines represent the averages of like count during different time periods, and illustrate the results of DiD models in Columns (3) and (4) of Table 4.2. The lighter red dash line during *VisUS* denotes the estimated average without the treatment effect ($FCRH \times Vis = -0.403, p < 0.05$) in Column (3) of Table 4.2.

T-NMR as the control group and compare it to F-CRH as the treatment group.

T-NMR VS F-NCRH. First, we conduct propensity score matching again with the same parameters as detailed in Section 4.4.2 to get matched T-NMR and F-NCRH groups with discarding 255 out of 26,271 tweets (0.97%). The biases of tweet- and author-related variables between the matched groups are all lower than 0.05. Second, we perform a parallel test between the treatment and control groups during the periods before and after the interventions. Prior to the interventions, both groups demonstrate parallel trends in misinformation diffusion in terms of retweet count and like count (Fig. S4.2(b)). Moreover, the unparalleled changes in T-NMR and F-NCRH groups during the two periods following the interventions are even more significant than those in F-CRH and F-NCRH groups (Fig. S4.2(a)). We then examine the virtual treatment effects using the DiD model. The coefficients of $TNMR \times Vis$ in Columns (1) and (3) of Table S4.4 are all statistically significantly negative ($coef. = -0.503, p < 0.05$ in Column (1); $coef. = -0.439, p < 0.05$ in Column (3)), just like those in Columns (1) and (3) of Table 4.2. Specifically, T-NMR tweets received 39.5% fewer retweets (95% CI: $[-59.5\%, -9.9\%]$) and 35.5% fewer likes (95% CI: $[-55.5\%, -6.6\%]$) compared to the baseline

Table 4.3: DiD regression results for retweet count and like count within F-CRH (treatment) and T-NMR (control) groups. Negative binomial regression models with robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

	Retweet Count		Like Count	
	(1) Pre-US	(2) US-World	(3) Pre-US	(4) US-World
<i>FCRH</i>	−0.289 (0.155)	−0.403*** (0.104)	−0.320* (0.149)	−0.327** (0.106)
<i>Vis</i>	−0.177 (0.132)	−0.665*** (0.083)	−0.130 (0.129)	−0.729*** (0.080)
<i>FCRH × Vis</i>	−0.184 (0.176)	0.029 (0.117)	−0.118 (0.177)	−0.022 (0.123)
<i>WordNum</i>	−0.060 (0.045)	0.047 (0.030)	−0.237*** (0.047)	−0.122*** (0.034)
<i>Media</i>	−0.079 (0.092)	0.286*** (0.057)	−0.078 (0.096)	0.127* (0.063)
<i>Hashtag</i>	−0.328 (0.226)	−0.488*** (0.114)	−0.530** (0.190)	−0.675*** (0.119)
<i>Mention</i>	−0.478*** (0.118)	−0.585*** (0.073)	−0.531*** (0.124)	−0.625*** (0.073)
<i>AcctAge</i>	−0.060 (0.041)	−0.087** (0.028)	−0.018 (0.042)	−0.092** (0.030)
<i>Verified</i>	−0.087 (0.111)	0.279*** (0.063)	−0.127 (0.111)	0.289*** (0.066)
<i>Followers</i>	0.251*** (0.030)	0.296*** (0.022)	0.285*** (0.029)	0.331*** (0.019)
<i>Followees</i>	−0.002 (0.064)	−0.016 (0.012)	0.042 (0.065)	0.002 (0.013)
Emotions	✓	✓	✓	✓
Topics	✓	✓	✓	✓
<i>Intercept</i>	9.163*** (0.164)	8.484*** (0.099)	10.919*** (0.156)	10.310*** (0.098)
Observations	1,716	6,463	1,716	6,463

before the first intervention and relative to the F-NCRH tweets. This indicates that with the expansion of Community Notes to the U.S., engagement with tweets in the T-NMR group also decrease even though they have no CRH notes. Also, the coefficients of $TNMR \times Vis$ are both not significant during the second intervention (Columns (2) and (4) of Table S4.4). With the F-NCRH group as the control group to compare, F-CRH and T-NMR groups show the same results regarding the effects of the interventions. This suggests that the decrease in the diffusion of misinformation may not be solely attributed to the targeted treatment of Community Notes on misleading tweets. Instead, it may be influenced by broader changes associated with the expansion of Community Notes, which could have impacted both misleading and non-misleading tweets in a similar manner.

F-CRH VS T-NMR. When comparing to the F-NCRH group, the F-CRH (misleading) and T-NMR (not misleading) groups show similar results. We then take T-NMR as the control group to examine the impact of the interventions on the engagement with misinformation within the F-CRH group. Based on the previous analysis in this study, it is expected that there would be no significant treatment effect on the F-CRH group when comparing the disparities between the F-CRH and T-NMR groups. To test this hypothesis, we first match the treatment and control groups through propensity score matching with the same parameters as described in Section 4.4.2. We discard 490 out of 8,989 (5.5%) tweets to get matched groups with all variable biases lower than 0.05. We then perform a

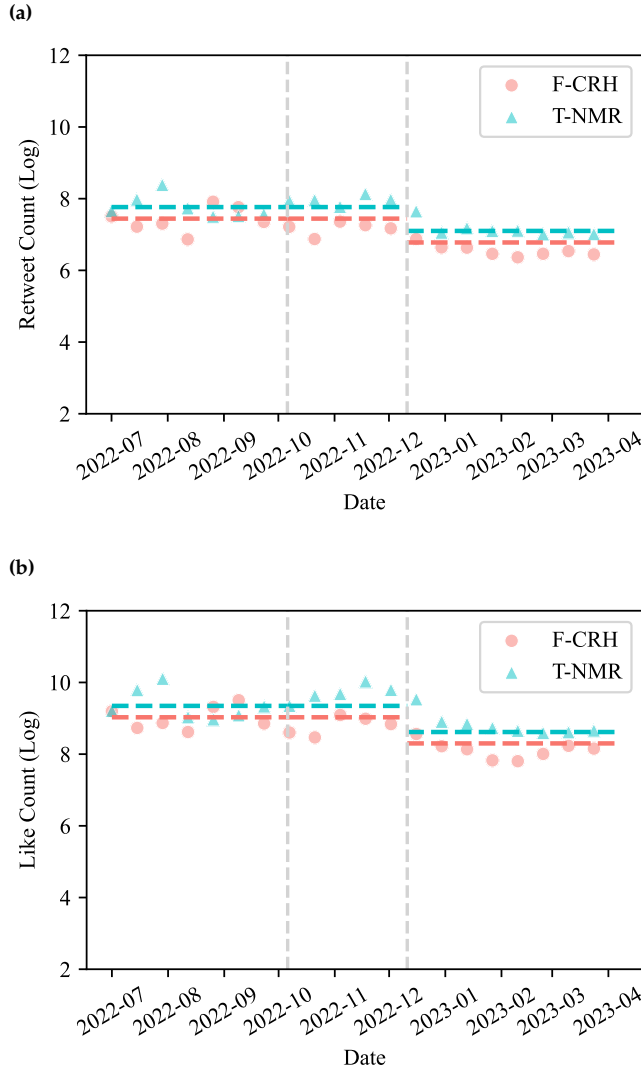


Figure 4.4: Two-week rolling averages of retweet count and like count in treatment (F-CRH) and control (T-NMR) groups from 2022-07 to 2023-04. The two vertical grey dash lines in each figure indicate two dates, October 6, 2022 (left) and December 11, 2022 (right), when Community Notes was expanded to the U.S. and the world respectively. (a) The points are 2-week rolling averages of retweet count. The horizontal dash lines represent the averages of retweet count during different time periods, and illustrate the results of DiD models in Columns (1) and (2) of Table 4.3. The two lines are always parallel over time, which corresponds to the statistically non-significant coefficient estimates of $FCRH \times Vis$ in Columns (1) and (2). (b) The points are 2-week rolling averages of like count. The horizontal dash lines represent the averages of like count during different time periods, and illustrate the results of DiD models in Columns (3) and (4) of Table 4.3. The two lines are always parallel over time, which corresponds to the statistically non-significant coefficient estimates of $FCRH \times Vis$ in Columns (3) and (4).

parallel test between the F-CRH and T-NMR groups across time periods. As expected, they have parallel trends regardless of the interventions (Fig. S4.2(c)). Using the DiD model, we further validate that there is no significant treatment effect for the two interventions (the coefficients of $FCRH \times Vis$ in the four columns of Table 4.3). As observed in Fig. 4.4, in addition to the parallel trends during the first intervention, the decreases in retweet count and like count within both F-CRH and T-NMR groups are no longer statistically significant (the coefficients of Vis in Columns (1) and (3) of Table 4.3). During the second intervention, the engagement with misleading and not misleading tweets significantly decreases in a parallel way, aligning with the previous findings in this study (the coefficients of Vis in Columns (2) and (4) of Table 4.3).

Additionally, we find that misleading tweets get not only fewer retweets (Fig. S4.3(a), $KS = 0.121, p < 0.001$), which supports earlier work conducted during the pilot phase of Community Notes [96], but also fewer likes than not misleading tweets (Fig. S4.3(b), $KS = 0.133, p < 0.001$). Furthermore, the T-NMR group consistently exhibits slightly higher retweet and like counts compared to the F-CRH group, regardless of the interventions (Fig. 4.4). The differences still exist under the control of other covariates (the coefficients of $FCRH$ in Table 4.3). The distributions

of retweet count and like count in misleading and not misleading tweets across different time periods are shown in Fig S4.4.

Consideration of the fact-checking delay. The above results do not provide evidence supporting that the roll-out of Community Notes specifically reduced engagement with misinformation in terms of retweet count and like count. Nevertheless, the effectiveness of Community Notes may be sensitive to fact-checking delays. We observed that the response time to note display is significantly longer than the impression half-life of source tweets (see Section 4.4.1), which suggests that notes were displayed after the most viral stages of diffusion. This delay has the potential to diminish the effect of Community Notes feature on reducing the aggregated engagement. Given this, we limit our analysis to a subgroup of tweets that received CRH notes in a relatively short delay and reevaluate the results. Notably, the shortest delay for note display in the dataset is 80.2 minutes, which is still longer than the half-life. Therefore, we lack sufficient samples to conduct the subgroup analysis around the half-life. Instead, we consider tweets that received CRH notes within 0.569 day (25% percentile) in the F-CRH treatment group, and the control group is still T-NMR. As reported in Table S4.5, the coefficients of $FCRH \times Vis$ in the four columns are all not statistically significant, which aligns with the results obtained without the limitation of the delay in note display.

Altogether, the results from the placebo tests imply that the roll-out of Community Notes did not lead to a statistically significant reduction in engagement with misinformation relative to truthful information. Instead, the diffusion of both true and false fact-checked tweets significantly decreased in parallel after the roll-out, which may be due to broader changes on the platform affecting both misleading and non-misleading tweets in a similar manner. This finding persists even when we restrict our analysis to tweets with a relatively short delay in note display. Furthermore, we obtain qualitatively identical findings when repeating the analysis without distinguishing between the two expansions of the Community Notes feature (see Table S4.6).

4.4.4 Analysis at the Note Display Level (RDD)

In the previous sections, we focused on different intervention periods and analyzed whether the introduction of the Community Notes feature and its roll-out to users in the U.S. and around the world has reduced engagement with misinformation on $\times$ /Twitter (*i.e.*, at the feature roll-out level). Next, we use our RDD model as a complementary method to study the effectiveness of the Community Notes feature at the note display level, that is, whether there is a reduction in engagement with misleading posts on which notes are displayed compared to those on which notes are not displayed.

To estimate our RDD model, we take *NoteScore* as the running variable to divide tweets into treatment and control groups with a cutoff point of 0.40. When fitting the RDD model, we only consider tweets posted after the roll-out intervention (roll-out period, as illustrated in Fig S4.1(b)). The treatment group, F-CRH, includes tweets with CRH notes that indicate they are potentially misleading and have helpfulness scores of no less than 0.40. The number of tweets in the F-CRH group is 2,133. Notably, if one F-CRH tweet has multiple CRH notes, the *NoteScore* is the average score of all its CRH notes. The control group, F-NCRH, includes tweets with non-CRH (NCRH) notes that indicate they are potentially misleading and

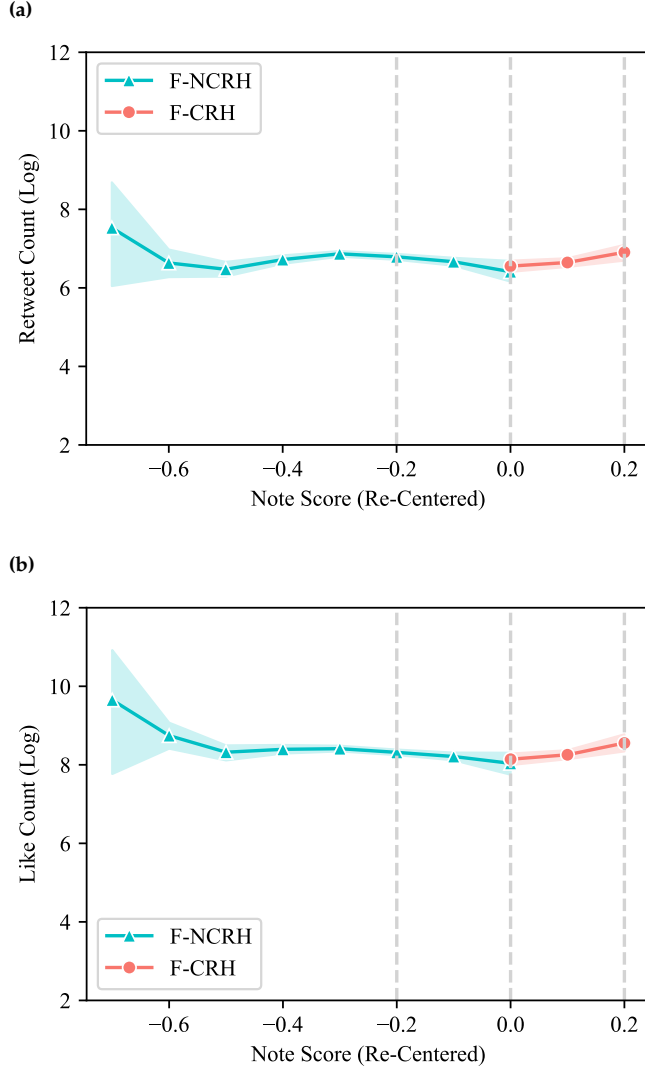


Figure 4.5: The averages of retweet count and like count across the note score points. Note scores are re-centered based on the cutoff point of 0.40. The shadow areas represent 95% Confidence Intervals. (a) The averages of retweet count across the note score points. There is no obvious discontinuity on the cutoff point. (b) The averages of like count across the note score points. There is no obvious discontinuity on the cutoff point.

have helpfulness scores of less than 0.40. Excluding tweets in the F-CRH group, F-NCRH has 9,671 unique tweets. Similarly, if one F-NCRH tweet has multiple NCRH notes, the *NoteScore* is the average score of all its NCRH notes.

We re-center *NoteScore*, and get the average retweet and like counts on different note score points. As depicted in Fig. 4.5, there is no clear discontinuity at the cutoff point. To test the discontinuity with the RDD model, we first set the window width as 0.20. The symmetrical distributions of retweet count and like count in the selected window are observed in Figs. 4.5(a) and (b). Next, we conduct propensity score matching with the same settings as detailed in Section 4.4.2. As a result, 325 out of 5,255 tweets (6.2%) are omitted, and all the variable biases between the two groups are lower than 0.05. Finally, we use the matched groups to perform RDD analysis. The coefficients of *Treated* in Columns (1) and (3) of Table 4.4 are not statistically significant. This indicates that there is no discontinuity at the cutoff point, which is consistent with the visual continuity at the cutoff point shown in Fig. 4.5. Moreover, despite limiting our analysis to tweets posted after the second intervention, during which the engagement with all fact-checked tweets shows a notable decrease (Fig 4.4), we still do not observe a significant effect of Community Notes

Table 4.4: RDD regression results for retweet count and like count within F-CRH (treatment) and F-NCRH (control) groups. Negative binomial regression models with robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

	Retweet Count		Like Count	
	(1) Roll-Out	(2) World	(3) Roll-Out	(4) World
<i>Treated</i>	0.157 (0.121)	0.114 (0.134)	0.180 (0.132)	0.044 (0.147)
<i>Delay</i>	0.018 (0.033)	−0.087 (0.053)	−0.020 (0.035)	−0.133* (0.053)
<i>NoteScore</i>	−0.871 (0.506)	−0.935 (0.562)	−0.303 (0.562)	−0.216 (0.623)
<i>WordNum</i>	0.038 (0.031)	0.076* (0.035)	−0.112** (0.035)	−0.070 (0.040)
<i>Media</i>	0.312*** (0.055)	0.423*** (0.062)	0.181** (0.060)	0.271*** (0.068)
<i>Hashtag</i>	−0.742*** (0.118)	−0.768*** (0.126)	−0.945*** (0.118)	−1.031*** (0.130)
<i>Mention</i>	−0.669*** (0.091)	−0.853*** (0.088)	−0.734*** (0.096)	−0.916*** (0.091)
<i>AcctAge</i>	−0.003 (0.028)	−0.012 (0.030)	0.013 (0.031)	0.001 (0.035)
<i>Verified</i>	0.142* (0.065)	0.258*** (0.073)	0.087 (0.070)	0.231** (0.082)
<i>Followers</i>	0.264*** (0.026)	0.282*** (0.040)	0.335*** (0.027)	0.326*** (0.033)
<i>Followees</i>	0.016 (0.023)	0.019 (0.024)	0.040* (0.020)	0.028 (0.019)
<i>Emotions</i>	✓	✓	✓	✓
<i>Topics</i>	✓	✓	✓	✓
<i>Intercept</i>	7.629*** (0.104)	7.369*** (0.116)	9.401*** (0.109)	9.163*** (0.127)
Observations	4,930	3,951	4,930	3,951

on reducing the engagement with misinformation (the coefficients of *Treated* in Columns (2) and (4) of Table 4.4). In addition, the coefficients of *Delay* in Table 4.4 are not significant or even negative (Column (4) of Table 4.4). This implies that the fact-checking speed did not have a significant effect in reducing the engagement with fact-checked tweets. Notably, this result is also consistent with our findings in the DiD analysis and previous work [96].

Altogether, the results of our RDD analysis provide no evidence to support the effectiveness of Community Notes in reducing engagement with misinformation in terms of retweet and like counts, which is consistent with our DiD analysis.

4.4.5 Confounding Analyses

To consolidate the findings of this study, we conduct additional analyses to incorporate potential confounding factors that are associated with the expansion of Community Notes on X/Twitter from multiple levels including the platform, Community Notes contributors, authors of tweets, and the tweets themselves.

Platform and contributors. From the platform level, Elon Musk concluded the acquisition of X/Twitter on October 27, 2022, which happened during the first expansion of Community Notes. After the acquisition,

Musk made some substantial changes on $\mathbb{X}$ /Twitter, which might affect the experience of users and the engagement of content on $\mathbb{X}$ [188]. This study uses a dummy variable, *Musk*, to control the possible effect of the acquisition of Elon Musk (= 1 if the tweet is posted after the acquisition, otherwise = 0). Moreover, we consider Month-Year fixed effects to further control for other potential time-related confounding factors. In addition, at the Community Notes feature level, with its expansion to the public, many new contributors are enrolled and write notes. Specifically, in the user enrollment dataset of this study, we find an enrollment of 48,928 new contributors (76% of all contributors) in Community Notes after its expansion to the world. Given the possible selection bias between old contributors in the pilot phase and new contributors enrolled after the expansion, we use another dummy variable, *OldWriter*, to control for the potential selection bias from contributors (= 1 if the tweet is noted by the contributors who had written notes during the pilot phase, otherwise = 0). With the two dummy variables, *Musk* and *OldWriter*, and Month-Year fixed effects included in the regressions, the coefficients and significance levels of $FCRH \times Vis$ are stable in Table S4.7, compared to Table 4.2. The coefficients of *Vis* and all the Month-Year fixed effects in Columns (1) and (3) of Table S4.7 are not statistically significant. This means that the changes in retweet count and like count before and after the expansion of Community Notes to the U.S., although slightly significant in Table 4.2, can be ignored. In Columns (2) and (4) of Table S4.7, the coefficients of *Vis* also become statistically non-significant, but the Month-Year fixed effects are all significantly negative during the months when Community Notes is open to the world. The statistically significant negativity of *Vis* transfers to the fine-grained Month-Year fixed effects in Columns (2) and (4) of Table S4.7 compared to Table 4.2. Particularly, the negative Month-Year fixed effects in 2023/02 (*coef.* = -1.186 , $p < 0.001$ in Column (2); *coef.* = -1.400 , $p < 0.001$ in Column (4)) and 2023/03 (*coef.* = -1.229 , $p < 0.001$ in Column (2); *coef.* = -1.389 , $p < 0.001$ in Column (4)) are stronger than those in 2023/01 (*coef.* = -0.785 , $p < 0.01$ in Column (2); *coef.* = -0.922 , $p < 0.001$ in Column (4)) and 2023/04 (*coef.* = -1.055 , $p < 0.001$ in Column (2); *coef.* = -1.355 , $p < 0.001$ in Column (4)). Month-Year fixed effects capture unobserved confounding factors that have significant impacts on the decrease of information diffusion and vary across different months. Additionally, the coefficients of *Musk* and *OldWriter* are not significant across all the columns in Table S4.7. This indicates that the acquisition and possible selection bias between new and old note authors have no substantial effect on the engagement with tweets noted by the Community Notes.

Furthermore, with the expansion of Community Notes, more and more $\mathbb{X}$ /Twitter users can transfer from retweeters to contributors who write or rate notes. This shift in user behavior has the potential to decrease the count of retweets or likes on individual tweets. To address the possible user transition, we redesign the dependent variables by incorporating the note count and rating count for each tweet. Specifically, the new retweet metric, *RetweetCountPlus*, denotes the combined count of retweets, notes, and ratings. Similarly, the new like metric, *LikeCountPlus*, represents the combined count of likes, notes, and ratings. In addition, we incorporate quote tweets (retweets with comments) into the retweets category and reexamine the regressions using the new metric *RetweetQuoteCount*. Using these new dependent variables, we regress F-CRH and F-NCRH groups again with the same independent variables. We find that the Month-Year fixed effects are still statistically significantly negative from 2023/01 to 2023/04 (see Table S4.8). Notably, for both retweet count and like count, the Month-Year effects from 2022/08 to 2022/12 are all

statistically non-significant (see Table S4.7 and Table S4.8). This suggests that the engagement with tweets in Community Notes has experienced no statistically significant change associated with unobserved time-specific factors before the second intervention. This finding is concordant with the horizontal trend of retweet count and like count from 2022/07 to 2022/12 in Fig. 4.3. The results for other key variables, such as $FCRH$ and $FCRH \times Vis$, are also robust.

Authors and tweets. In this study, we use PSM to balance the samples between treatment and control groups based on the characteristics from the author profile and tweet content. However, it is important to note that the differences in the samples across different time periods may be not fully balanced. We further consider the potential influence of unbalanced factors during different time intervals. Based on matched F-CRH and F-NCRH groups, we further match tweets posted before (during *Pre1*) and after (during *VisUS*) the expansion of Community Notes feature to the U.S. on October 6, 2022 [168]. 1,214 out of 5,028 (24.1%) tweets are omitted, and the mean bias of the variables is 0.017. Similarly, we match tweets posted before (during *VisUS*) and after (during *VisWorld*) the expansion of Community Notes feature to the whole world. 5,836 out of 11,842 (49.3%) tweets are omitted, and the mean bias of the variables is 0.014. We do the same regression analysis as presented in Table S4.7 using the newly matched tweets, and the results remain consistent and stable (see Table S4.9).

Although we mitigate the differences in author profiles among different groups in terms of account age, verified status, follower count, and following count, there are still some unobserved factors that are associated with authors and can affect engagement with online information. To control the unobserved factors related to the authors, we only consider the tweets posted by the same users over different time periods. Additionally, we also need to consider the change in the engagement with content at the platform level to analyze whether the engagement with tweets with notes (included in the Community Notes) is different from those without notes (out of Community Notes).

To do so, we use F-CRH as the treatment group, which includes 3,387 tweets posted from 2,153 unique authors. First, we collect 374,601 new tweets posted by the same authors in the F-CRH group during the three time periods, namely, *Pre1*, *VisUS*, and *VisWorld* (from 01-07-2022 to 06-04-2023). The new tweets without notes represent the group OCN (Out of Community Notes) and undergo the same preprocessing steps as described in Section 4.3.2. Second, during *Pre1*, 210 authors are included in the F-CRH group. We use tweets posted by these authors in F-CRH and OCN groups to analyze the first intervention between *Pre1* and *VisUS*. Specifically, there are 313 tweets in F-CRH and 11,427 tweets in OCN. Given the substantial disparity in tweet counts between OCN and F-CRH groups, we conduct one-to-one matching with a narrower caliper of 0.0001 to ensure precise matches. As a result, 269 out of 313 tweets (85.9%) in the F-CRH group have corresponding matches in the OCN group. Third, during *VisUS*, 515 authors are included in the F-CRH group. Tweets from these authors in F-CRH and OCN groups are used to analyze the second intervention between *VisUS* and *VisWorld*. We perform one-to-one matching again for 976 tweets in the F-CRH group, and find 937 corresponding matches (96%) from 83,505 tweets in the OCN group. Using matched tweets in both groups, we find that the coefficients of $FCRH \times Vis$ in Table S4.10 are not significant across all columns. The extra reduction in engagement within the F-CRH group is no longer significant compared to the baseline before the interventions

and relative to the OCN group. This further fails to support that the implementation of the Community Notes feature on $\mathbb{X}$ /Twitter had a statistically significant impact on the aggregated retweet count and like count in misleading tweets, thereby supporting the main claim of this paper. Moreover, in Table S4.10, the coefficients of *Vis* and Month-Year fixed effects are not statistically significant, even in Columns (2) and (4). This indicates that the engagement of content on $\mathbb{X}$ /Twitter shows no significant variation for the same authors across different time intervals. It also implies that the decreases in retweet volume and likes are attributed to the inclusion of fact-checked tweets from authors who were newly included in each time period. Additionally, the results remain robust even under the shock of Musk's takeover.

In addition to the retweet count and like count, the engagement rate, which represents the ratio of retweets (retweet rate) or likes (like rate) to the total views, is another important metric for assessing the performance of Community Notes. This metric is utilized by $\mathbb{X}$ /Twitter itself to measure the effectiveness of the feature [67]. Starting from December 15, 2022, $\mathbb{X}$ /Twitter has released a new public metric, impression count, to show the total number of times a Tweet has been viewed [180]. Because the impression count is only available after the implementation of Community Notes on the whole platform of $\mathbb{X}$ /Twitter, we cannot conduct the same regression analysis for retweet rate (like rate) as we do for retweet count (like count). Accordingly, we employ a binomial regression to investigate the association between Community Notes and engagement rate. The findings in Table S4.10 demonstrate that controlling tweets from the same authors in different time periods can effectively mitigate the influence of confounding factors. In the analysis of the engagement rate, we focus on tweets from 515 authors who are included in the Community Notes during *VisUS*. Specifically, there are 328 tweets in the F-CRH group and 62,994 tweets in the OCN group after the inclusion of impression count as a public metric. Through one-to-one matching, we identify corresponding matches for 306 out of 328 tweets (93.3%) in F-CRH (treatment group) from OCN (control group). We assume that the number of retweets (likes) for a given tweet i follows a binomial distribution that is specified as

$$RetweetCount_i(LikeCount_i) \sim \text{binomial}(ImpressionCount_i, \theta_i), \quad (4.3)$$

where $\theta_i \in [0, 1]$ represents the probability that a user will retweet or like tweet i , and corresponds to the retweet (like) rate. We then use a *logit* link function and specify the binomial regression model as:

$$\text{logit}(\theta_i) = \beta_0 + \beta_1 Treated_i + \alpha' x_i, \quad (4.4)$$

where $Treated_i = 1$ if tweet i is in the treatment group (*i.e.*, F-CRH group), otherwise = 0. β_0 is the intercept, and x_i denotes all the control variables described in Section 4.3.2.

As shown in Table 4.5, the coefficients of *FCRH* in Columns (1) and (3) are statistically non-significant. It implies that the retweet rate and like rate are not significantly associated with Community Notes. This finding is robust under the control of Month-Year fixed effects in Columns (2) and (4). Additionally, we observe statistically significant negative coefficients for the variable *Media* across all the columns in Table 4.5. Considering the positive effects of *Media* on retweet and like counts in the previous findings, as shown in Table 4.2, we can interpret that tweets with media elements tend to have higher aggregated retweet (like) counts. However, it is important to note that these tweets also exhibit lower retweet (like)

Table 4.5: Regression results for retweet rate and like rate within F-CRH (treatment) and OCN (control) groups. Binomial regression models with robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

	Retweet Rate		Like Rate	
	(1)	(2)	(3)	(4)
<i>FCRH</i>	−0.014 (0.249)	−0.031 (0.263)	−0.388 (0.256)	−0.433 (0.253)
<i>WordNum</i>	0.228** (0.083)	0.236** (0.081)	−0.058 (0.076)	−0.063 (0.075)
<i>Media</i>	−0.616*** (0.136)	−0.602*** (0.147)	−0.529*** (0.122)	−0.550*** (0.136)
<i>Hashtag</i>	−0.586 (0.401)	−0.480 (0.336)	−0.454 (0.303)	−0.268 (0.251)
<i>Mention</i>	0.725 (0.404)	0.724 (0.404)	0.355 (0.329)	0.400 (0.331)
<i>AcctAge</i>	−0.044 (0.059)	−0.040 (0.063)	−0.161** (0.053)	−0.161** (0.057)
<i>Verified</i>	−0.184 (0.250)	−0.183 (0.264)	−0.096 (0.219)	−0.078 (0.232)
<i>Followers</i>	−0.103 (0.053)	−0.103 (0.053)	−0.078** (0.027)	−0.085** (0.027)
<i>Followees</i>	0.060 (0.062)	0.062 (0.063)	0.000 (0.049)	0.012 (0.051)
<i>MonthYear</i>	✗	✓	✗	✓
<i>Emotions</i>	✓	✓	✓	✓
<i>Topics</i>	✓	✓	✓	✓
<i>Intercept</i>	−6.285*** (0.347)	−6.260*** (0.404)	−4.343*** (0.341)	−4.120*** (0.382)
Observations	612	612	612	612

rates when compared to tweets without media elements.

In sum, we conducted a thorough analysis on potential confounding factors stemming from the platform, contributors, tweet authors, and tweets themselves. The findings consistently fail to provide significant evidence supporting that the roll-out of Community Notes led to a reduction in engagement with misinformation in terms of aggregated retweet counts and likes. Furthermore, our analysis implies that the observed reduction in retweets and likes after the second intervention for both misleading and not misleading posts can be partially attributed to differences in the fact-checking targets, *i.e.*, the incorporation of fact-checked tweets from authors who were newly included when the Community Notes feature was rolled out to the world.

4.5 Discussion

The rise of social media provides unprecedented opportunities for individuals to share information and engage in public discourse. However, it has also created an environment where users can engage with misinformation easily, leading to negative impacts on individuals, society, and democracy as a whole. As such, engagement with online misinformation has become an important issue, and many interventions have been developed to address this problem. Evaluating the effectiveness of these interventions is crucial to ensure that efforts are directed towards the most promising strategies, and to inform future policy decisions. To our best knowledge, this study is the first to rigorously examine the

effectiveness of crowdsourced fact-checking in reducing engagement with misinformation in the real world.

We use a longitudinal crowdsourced dataset of fact-checks from Community Notes to investigate whether this new feature effectively reduces engagement with misinformation on X/Twitter (as measured by retweet and like counts). Despite a significant increase in the number of notes and fact-checked tweets, we find that the expansion of Community Notes to the U.S. and the world might not have had a significant effect on reducing engagement with misinformation. This is confirmed by both DiD and RDD models, as well as comprehensive confounding analysis. Instead, our study reveals that the diffusion of both true and false fact-checked tweets significantly decreased in parallel when Community Notes became available to the world. Additionally, our results show that misleading tweets have lower virality compared to non-misleading tweets in terms of retweet and like counts, regardless of the expansion of the Community Notes feature.

4.5.1 Research Implications

Our analysis provides several important research implications. As our main contribution, we demonstrate that the roll-out of Community Notes had no statistically significant effect on reducing engagement with misinformation on X/Twitter in terms of retweet count and like count. This highlights the significance of utilizing real-world data for the evaluation of misinformation interventions [189].

At a first glance, our findings seem to be inconsistent with the results of an A/B test carried out by X during the pilot phase of Community Notes [67]. X found that users exposed to note annotations on tweets were 25–34% less likely to like or retweet them compared to the control group. From this perspective, misleading tweets with CRH notes are expected to have lower engagement, compared to misleading tweets without CRH notes. However, this difference in engagement is not observed in our study. Importantly, however, this inconsistency does not necessarily mean that Community Notes cannot inform users and influence their engagement behaviors. Rather, a plausible explanation might be that the delay in note display to users may partially explain the insufficiency of Community Notes in reducing engagement with misinformation. While our subgroup analysis regarding different fact-checking delays fails to find supporting evidence for this notion, this may be because the smallest observed fact-checking delays in our data are still not short enough compared to the spread of source tweets. A recent study shows that about 95% of tweets have no relevant impressions after two days, and it only takes about 79.5 minutes before half of the overall impressions are created for a tweet [180]. Similarly, research on Weibo, an X/Twitter-like service in China, suggests that approximately 80% of reshares occur within the first two days of a post [10]. Nevertheless, the results of our study reveal that, on average, fact-checking CRH notes require more than two days to appear directly on the corresponding source tweets, with the shortest response time observed being 80.2 minutes. This delay thus limits their ability to inform users at the early and most viral stage of misinformation diffusion.

Our current study primarily focuses on evaluating the aggregated engagement metrics (*i.e.*, retweet count and like count) before and after the expansion of the Community Notes feature, rather than specifically examining the engagement timeline in each misleading tweet before and

after the addition of notes. Given this limitation, we further consider potential confounding factors from multiple levels, including platform, contributors, authors, and tweets, to support our findings. Nonetheless, we also highlight the need for more detailed investigations to scrutinize the timelines and specific patterns of engagement change. Such fine-grained studies can provide deeper insights into the effects of Community Notes on user engagement. Theoretically, fact-checking annotations can reduce people's belief in misinformation [67, 190], but also may recall the backfire effect, which inadvertently strengthens misconceptions [1]. Moreover, previous research has indicated that users still engage with online misleading posts they distrust [15, 191, 192]. Additionally, despite existing deficits regarding critical reasoning and partisan bias in discerning falsehood, social media users also share information that challenges their own political ideologies [193]. Since engaging with misinformation is part of broader complex social reactions [193, 194], future research should analyze the role of such patterns within crowd-based fact-checking in real-world social media environments.

We further observe that the engagement with misleading and not misleading fact-checked tweets decreased in parallel after Community Notes was expanded to the world. Several factors may contribute to this pattern. On the platform side, X/Twitter claims that they do not take any enforcement of other rules apart from the display of CRH notes on fact-checked tweets via Community Notes. Additionally, Musk's takeover and other time-specific confounding factors (such as real-world events) cannot explain the decline (see Table S4.7). From the aspect of community contributors, we find an enrollment of 48,927 new contributors (76% of all contributors) in Community Notes after its expansion to the world, which is more than 24 times the number of enrolled contributors in the expansion limited to the U.S. With such a massive increase in the size of the community, unobserved changes may occur in the profiles of community contributors. However, differences between old contributors and new contributors are not statistically significant (see Table S4.7). Finally, when only comparing tweets from the same authors during different time periods, we find that the decline disappears, which means the decreases in retweet and like counts can be attributed to the inclusion of fact-checked tweets from authors who were newly included in each time period (see Table S4.10). This also implies that the contributors, for both old and new enrolled ones, may focus on a wider range of tweets and not only the most viral ones after the pilot phase of Community Notes ended.

Ultimately, we find that community fact-checked misleading tweets consistently receive fewer retweets and likes than non-misleading tweets, regardless of whether the Community Notes program was expanded or not. A recent analysis of misinformation diffusion, which is based on the tweets fact-checked by Community Notes during its early pilot phase, also finds this phenomenon [96]. Notably, this finding is contrary to the previous research that examines the spread of misinformation fact-checked by experts from third-party organizations [14]. One potential explanation for this discrepancy is coming from the sample selection, *i.e.*, that community fact-checkers may prioritize posts from influential accounts [96, 150, 182]. Additionally, our study also finds a positive correlation between verified accounts with a higher number of followers and receiving more notes and ratings. Nevertheless, the higher level of diffusion in not misleading tweets still exists under the control of the variables that denote the influence of authors (the coefficients of *FCRH* in Table 4.3), which calls for more detailed explanations.

4.5.2 Practical Implications

This study also provides important practical implications for harnessing the “wisdom of crowd” as a countermeasure against misinformation on social media platforms. First, the significant increase of individual fact-checkers in Community Notes after its implementation on X/Twitter makes it feasible to fact-check online information at large scale. Moreover, through active individual fact-checkers on social media, specific pieces of misinformation that come from highly influential accounts or raise significant concerns can be checked directly at their original places.

Second, while crowdsourced fact-checking, such as Community Notes, can provide a scalable way to detect and label misinformation [46, 96], they may not lead to significant reductions in the engagement with misinformation. The lack of a significant effect of Community Notes on mitigating overall engagement with misinformation should remind social media platforms to reconsider their crowdsourced fact-checking strategies. Reliable verdicts require sufficient assessments from different perspectives, which means that corresponding pieces of misinformation must first gain attention and become widespread to receive enough ratings. To address this post-diffusion dilemma with crowdsourced fact-checking, social media platforms could implement a pre-selection mechanism to filter tweets with a high probability of being misleading at the beginning of their diffusion and then selectively push them to community fact-checkers. By incentivizing contributors to act promptly, such a pre-selection mechanism could reduce delays in obtaining fact-checking verdicts. Notably, while Community Notes may not have directly reduced the overall engagement with misinformation, it may still provide a valuable feature to keep users well-informed with helpful context.

Third, our study also highlights the importance of measuring the impact of fact-checking interventions using rigorous methods and real-world data, rather than relying solely on controlled experiments or simulations. By analyzing large-scale datasets of user behaviors and social interactions, researchers can gain a better understanding of the complex dynamics of misinformation propagation and evaluate the effectiveness of various interventions.

Finally, our study also underscores the importance of improving media literacy among users. Social reactions, such as retweet and like, are the essential components that make social media “social” [194]. There are complex interaction mechanisms among online users. While fact-checking can be an effective tool for identifying and flagging false claims, it may not be sufficient to reduce engagement with misinformation. Therefore, social media platforms, governments, and public organizations should work together to promote critical thinking, digital literacy, and transparency in online communication [166].

4.5.3 Limitations and Future Research

There are several limitations in our study. First, because the dataset is mainly in English and the program is well-established in the U.S., we conduct the research with tweets only in English. Multilingual evaluation should be examined with the establishment of Community Notes to get a more comprehensive understanding. Second, we focus on the effect of Community Notes on the engagement with misleading tweets and use the T-NMR group as a mediating group indicating non-misleading information. However, tweets in the T-NMR group also have the potential

to be misleading, because of general misconception or manipulation [50, 51, 96]. To mitigate this concern, we implement rigorous selection criteria to ensure that only tweets with a high likelihood of being non-misleading are included in the T-NMR group. Relatedly, F-NCRH tweets serve as the control group for comparison with the F-CRH group. However, certain F-NCRH notes may not be rated as helpful because the corresponding source tweets are not misleading (*e.g.*, due to misuse of the fact-checking feature). To account for this, we exclude T-NMR tweets from F-NCRH tweets, specifically those containing both T-NMR notes and F-NCRH notes, and use propensity score matching to filter the F-NCRH group to enhance its comparability with the F-CRH group to the largest extent possible. Third, 14% of the total tweets fact-checked via Community Notes are unavailable via the API (see Section 4.3.1) and thus cannot be included in the evaluation. Some of these unavailable tweets might be misleading and removed by the authors when the CRH notes appeared. Additionally, some particularly egregious misinformation might have been removed by the platform (*i.e.*, X/Twitter). While the retrieval rate for tweets with CRH notes is similar to the overall rate for all tweets, this could potentially still lead to an underestimation of the effectiveness of Community Notes. Fourth, our study is limited to X/Twitter, and the findings may not be generalizable to other social media platforms with different user demographics and content characteristics. Finally, given the challenges of accessing detailed engagement timelines from X/Twitter, we had to rely on aggregated retweet count and like count as metrics to analyze the effectiveness of Community Notes. Future research should try to further explore the effect of Community Notes and other misinformation interventions with more fine-grained social media data. This would provide a more comprehensive understanding of the effects and mechanisms underlying these interventions, potentially leading to more effective strategies in reducing engagement with misinformation on social media platforms.

4.6 Conclusion

In this paper, we conduct a large-scale empirical study to evaluate the effectiveness of X's Community Notes feature in reducing engagement with misinformation on X/Twitter. Based on the results from both DiD and RDD models, we find no evidence that the introduction of Community Notes and its roll-out to users in the U.S. and around the world led to a significant reduction in user engagement with misleading tweets. Our findings suggest that the response time of Community Notes may be not fast enough to intervene in the engagement with misinformation during early and highly viral stages of diffusion. Taken together, our work emphasizes the importance of evaluating fact-checking interventions in the field and provides important implications for improving crowdsourced fact-checking strategies on social media platforms.

Effects of Visibility on Overall Engagement

5

Abstract: In June 2024, X/Twitter changed likes’ visibility from public to private, offering a rare, platform-level opportunity to study how the visibility of engagement signals affects users’ behavior. Here, we investigate whether hiding liker identities increases the number of likes received by *high-reputational-risk content*, content for which public endorsement may carry high social or reputational costs due to its topic (e.g., politics) or the account context in which it appears (e.g., partisan accounts). To this end, we conduct two complementary studies: 1) a Difference-in-Differences analysis of engagement with 154,122 posts by 1068 accounts before and after the policy change. 2) a within-subject survey experiment with 203 X users on participants’ self-reported willingness to like different kinds of content. We find no detectable platform-level increase in likes for high-reputational-risk content (Study 1). This finding is robust for both between-group comparison of high- versus low-reputational-risk accounts and within-group comparison across engagement types (*i.e.*, likes vs. reposts). Additionally, while participants in the survey experiment report modest increases in willingness to like high-reputational-risk content under private versus public visibility, these increases do not lead to significant changes in the group-level average likelihood of liking posts (Study 2). Taken together, our results suggest that hiding likes produces a limited behavioral response at the platform level. This may be caused by a gap between user intention and behavior, or by engagement driven by a narrow set of high-usage or automated accounts.

- 5.1 Introduction 89
- 5.2 Related Work 91
 - 5.2.1 Social Media Affordances 91
 - 5.2.2 Content Moderation and Self-Censorship 92
- 5.3 Study 1: Observational Analysis 93
 - 5.3.1 Methods 93
 - 5.3.2 Empirical Results 95
- 5.4 Study 2: User Study 98
 - 5.4.1 Experiment Setup 98
 - 5.4.2 Experimental Results 98
- 5.5 Discussion 101

Preprint

Y. Chuai, M. H. Ribeiro, G. Lenzini, and N. Pröllochs, ‘When “likers” go private: Engagement with reputationally risky content on X,’ *ArXiv*, 2026.

5.1 Introduction

Social media platforms scaffold public expression through interface *affordances* that make some actions easy, legible, and rewarding while rendering others costly or invisible [195–197]. Among these, the *visibility* of engagement signals is central: visible cues act as social proof and reputation broadcasts, shaping both audiences’ inferences and authors’ self-presentation strategies [198].

In June 2024, X (formerly Twitter) implemented a major design change by making liker identities private: users can still like posts, and authors can see who liked them, but third parties can no longer inspect others’ likes (see Fig. 5.1). This policy directly reconfigures the platform’s visibility affordance for liking without altering the act of posting, offering a rare, platform-scale case to study how privacy of engagement changes behavior.

But why might hiding liker identities matter? First, reputational and normative pressures can deter the visible endorsement of sensitive content, leading to self-censorship [197, 199]. Second, visible engagement

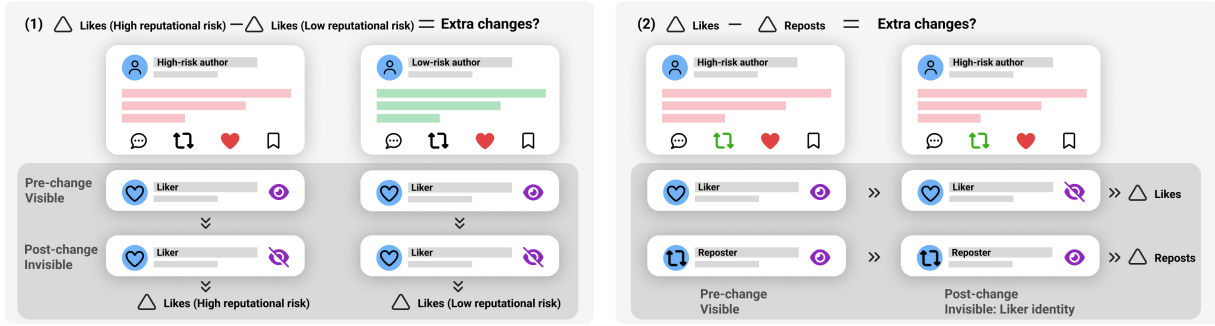


Figure 5.1: Research overview. $\times$ made public likes private on June 12, 2024. After the change, the identities of likers are visible only to themselves and the post authors. To evaluate the behavioral impact of this change, we estimate (1) the additional changes in likes received by high-reputational-risk accounts after the design change, relative to the pre-change period and compared to low-reputational-risk accounts (between-group); and (2) the additional changes in likes to high-reputational-risk accounts after the design change, relative to the pre-change period and compared to reposts on the same accounts (within-group).

metrics amplify bandwagon dynamics and moralized responses [74, 77]. If reputational risk is a primary brake on liking, privatizing the liker’s identity should *increase* engagement with high-reputational-risk content (content that users would feel social pressure not to engage in public; *e.g.*, partisan or lewd posts), while leaving low-reputational-risk content largely unchanged.

Beyond social dynamics, like visibility also matters for the platform’s algorithmic layer. Engagement signals such as likes, reposts, and replies are integral inputs to ranking and recommendation systems that determine which posts are surfaced to users [200, 201]. Publicly visible likes historically functioned as both a *social* and an *algorithmic* feedback loop: they indicated popularity to users while signaling content quality to the platform. By privatizing liker identities, $\times$ decouples these two layers, reducing the social signaling function while leaving the algorithmic weight of likes intact. This shift could alter how engagement propagates through the feed: if users become marginally more willing to like controversial content, the ranking system may amplify it even without visible social proof. Conversely, if liking remains concentrated among a small subset of users or automated accounts, the algorithm’s exposure dynamics may stay largely unchanged [202, 203]. Understanding whether this design change modifies not only users’ visible behavior to controversial content but also the algorithmic distribution of content is key to assessing its civic and informational consequences.

Present work. Here, we use $\times$ ’s policy change as a quasi-experiment and ask:

- **RQ1:** Did making likes private change the number of likes received by high-reputational-risk accounts relative to low-reputational-risk accounts?
- **RQ2:** Did likes change relative to unaffected engagement types (*e.g.*, reposts) for high-reputational-risk content?
- **RQ3:** Do users say they would like more when likes are private, especially for controversial topics?

We answer RQs 1–2 with an observational analysis of 154,122 posts from 1068 accounts spanning eight weeks pre-change and four weeks post-change. We estimate effects with Difference-in-Differences (DiD) approach across two complementary designs: (i) between groups (high-vs. low-reputational-risk accounts) and (ii) within posts (likes vs. reposts), and probe pre-trends and inference robustness with HonestDiD and equivalence test. To address RQ3 and mechanism plausibility, we run a survey experiment in which participants rate their likelihood of liking

posts on high-reputation-risk and low-reputational-risk topics under public vs. private visibility.

Findings. Across both DiD designs, weekly and aggregated effects are statistically indistinguishable from zero: the privacy change did not yield a detectable platform-level increase in likes to high-reputational-risk content, nor a shift in likes relative to reposts. In the survey, participants *report* modest within-person increases in willingness to like controversial content under private visibility, but group means remain similar across visibility conditions. Together, our results suggest a limited behavioral response at scale despite stated preferences, which is consistent with an intention–behavior gap and the concentration of engagement among specialized, high-usage, or automated accounts.

Contributions. Our contributions are three-fold:

1. A causal estimate of how privatizing liker identities affects engagement with high-reputational-risk content, using two DiD strategies with robustness checks.
2. A mixed-methods triangulation that contrasts platform-level behavior with user-level stated preferences to illuminate mechanisms.
3. Design implications for visibility interventions: hiding *who* liked may be insufficient to meaningfully shift sensitive engagement when other incentive and distributional forces dominate.

We discuss implications for moderation and measurements on social media platforms: visibility changes that target reputational risk may require complementary levers (*e.g.*, distribution, feedback timing, or audience segmentation) to move behavior in consequential ways.

5.2 Related Work

We briefly review previous work on social media affordances, content moderation, and self-censorship.

5.2.1 Social Media Affordances

Social media affordances refer to the perceived, actual, and imagined properties of platforms that emerge through the interplay of technological, social, and contextual factors, shaping and constraining how users engage with them [195, 196]. These affordances are generally conceptualized along two levels: high-level affordances capturing the abstract capabilities of platforms, and low-level affordances, embedded in the material design of specific features, such as buttons, screens, and interactive elements [195, 204].

The four-item high-level taxonomy of affordances—visibility, editability, persistence, association—applies to a range of social media platforms and covers the most significant and controversial aspects of the use of social media [196]. Visibility refers to the extent to which information, behaviors, and connections are observable to others within the platform. Editability captures users’ ability to craft, revise, and refine content before or after publication, allowing for greater control over self-presentation. Persistence denotes the durability of digital traces over time, as content remains accessible and searchable long after initial creation. Finally, association describes the capacity of social media to link individuals, content, and communities through features such as tagging, following, or connecting.

To better understand self-presentation on social media platforms, prior research proposed an affordance-based framework, which includes three main components related to the self (presentation flexibility, content persistence, identity persistence), other actors (content association, feedback directness), and the audience (audience transparency and visibility control) [197].

The affordances of social media platforms are constrained by engagement features, *i.e.*, low-level affordances [195]. Likes, comments, and shares are three main engagement features across many social media platforms such as X (formerly Twitter), Facebook, and TikTok. The value affordances of these engagement features are summarized by previous research, and each feature is associated with specific trade-offs [205]. Specifically, the Like feature encourages recognition and positivity but undermines authenticity; the Comment/Reply feature fosters feedback and diversity but invites conflict; and the Sharing feature strengthens togetherness but can pressure users and threaten privacy. For instance, by rewarding specific types of posts through “likes,” users not only encourage others to produce similar content but also shape their own content exposure, as platform algorithms curate their feeds based on prior engagement patterns [200, 201].

5.2.2 Content Moderation and Self-Censorship

Social media platforms provide users with considerable flexibility for expression, yet they enable the creation and dissemination of controversial or harmful content, such as misinformation, hate speech, and harassment [14, 206–211]. Consequently, platforms face growing pressure to design effective moderation mechanisms that mitigate harm while safeguarding freedom of expression. Existing content moderation interventions are largely designed to act upon or reshape the affordances of social media. For instance, the use of warning labels, annotations, down-ranking, and visibility reduction directly targets the visibility affordance by limiting the exposure of problematic or misleading content [44, 52, 74, 76, 142, 212, 213]. Similarly, the editing or removal of posts and the deplatforming of user accounts are tied to the editability and persistence affordances, as they determine whether and how content can be altered, deleted, or remain accessible over time [185, 214–222].

Notably, from the user perspective, visibility is a consequential affordance, constituting the social dimension of online interaction. High visibility can promote exposure to a broad audience but also intensifies social evaluation, leading users to self-censor content they perceive as controversial or reputationally risky [196, 197]. *Self-censorship refers to the deliberate act of withholding or altering expression to avoid potential backlash or sanctions* [199]. This behavior is reinforced by platform feedback systems and community-driven norms, such as post guidance [223], community notes that flag misleading posts [74, 76, 213], and collective reactions of moral outrage [77]. For example, the display of community notes can increase the odds that authors delete their misleading posts by 103.4% [76], demonstrating how visibility-based interventions can shape user behavior.

Users’ perceptions of being observed also influence their willingness to disclose personal/sensitive information. Heightened surveillance awareness amplifies privacy concerns and reduces the perceived value of sharing, prompting protective behaviors such as self-censorship [224, 225]. Additionally, the perceived platform audience composition plays a

crucial role in disclosure decisions involving sensitive content [226]. On X, users strategically navigate visibility boundaries by toggling between public and protected account settings to manage audience exposure and personal disclosure [227, 228].

From the audience or observer perspective, anonymous sharing is a popular choice, especially for controversial content, suggesting that users rely on anonymity to mitigate reputational risks [229]. Social media anonymity can also enhance moral courage, as it reduces the perceived risk of adverse consequences for taking moral actions [230]. Beyond anonymity, the visibility of engagement metrics significantly influences how users interact with content. For example, the public visibility of likes makes this form of engagement more susceptible to normative social influence than less visible behaviors, such as clicking or viewing [231, 232]. Users also strategically manage their public personas through visible engagement cues, for instance, withholding likes on politically sensitive content to avoid potential reputational consequences [198].

Research gap. On X, recent platform change—specifically, the shift from public to private likes—represents a significant reconfiguration of the visibility affordance [233, 234]: users can see posts they have liked (but others cannot); like count and other metrics for users’ own posts will still show up under notifications; users no longer see who liked someone else’s post; and post authors can see who liked their posts. This design alteration has the potential to reshape patterns of engagement by reducing the normative and reputational pressures associated with visible liking behavior to the controversial content. Examining how user interactions change following this policy shift is therefore crucial for understanding how visibility mediates expression and self-censorship in contemporary social media environments, which presents the aim of this study.

5.3 Study 1: Observational Analysis

We conduct an empirical analysis to examine whether making liker identities private on X produces measurable shifts in the number of likes to high-reputational-risk accounts that are more likely to trigger users’ self-censorship than low-reputational-risk accounts.

5.3.1 Methods

Data Collection

X made “likes” private on June 12, 2024. In Study 1, we collect posts from identified accounts between eight weeks before (−8 weeks) and four weeks after the date (+4).

Selecting X accounts. We consider X accounts belonging to categories that may or may not be affected by changes in self-censoring behavior following X’s feature change, and classify them into *high-reputational-risk* and *low-reputational-risk* groups [229, 235]. This distinction is not intended to judge the nature of the content itself, but rather to capture that users may hesitate to publicly engage with certain accounts due to perceived social or reputational pressures (*e.g.*, interactions with politically sensitive accounts). We consider the following *high-reputational-risk* accounts:

Table 5.1: Overview of accounts, posts, and their associated engagement of likes and reposts.

	(1) #Accounts	(2) #Posts	(3) #Likes	(4) #Reposts
High-reputational-risk				
Extremist	31	8,521	20,314,179	4,591,615
OnlyFans	144	34,926	16,430,401	1,636,402
Political influencer	66	30,444	187,296,366	33,332,385
Politician	566	66,417	44,854,643	9,963,139
Low-reputational-risk				
Actor	53	1,399	9,909,943	1,166,771
Gamer	72	2,749	7,426,351	245,959
Singer	69	2,246	35,759,453	5,179,803
Tech influencer	67	7,420	4,258,423	225,137
Total	1,068	154,122	326,249,759	56,341,211

- *SPLC extremist accounts (N=44)*. Accounts belonging to individuals the Southern Poverty Law Center (SPLC) listed as extremists [236]. The list includes 144 extremists. We manually search for their accounts on X and successfully find 44 active ones throughout the observation period.
- *Politicians (N=632)*. We obtain a dataset of politicians' X handles from Wikidata [237]. Specifically, we consider former U.S. Presidents as well as members of the 118th U.S. Congress (both Senate and House of Representatives). We supplement the political ideology scores for the politicians [238].
- *Political influencers on X (N=69)*. AllSides X Influencer Bias Chart (Version 1.1, see Fig. S5.1) includes bias ratings for the most influential content creators on X [239]. Based on this chart, we collect their linked account handles on X.
- *OnlyFans accounts (N=229)*. We consider accounts associated with sex workers on OnlyFans. The official OnlyFans handle regularly promotes its member accounts. Given this, we first collect all the posts created by the OnlyFans official account during 2023 and extract the member accounts mentioned in the posts.

In addition, we consider the following *low-reputational-risk* accounts:

- *Tech influencers (N=71)*. We select tech influencers with active X accounts based on a 2025 list of top tech influencers [240].
- *Gamers (N=100)/Singers(N=100)/Actors (N=100)*. We again query Wikidata to search for gamers, singers, and actors. Subsequently, we select the top 100 professionals with the highest number of followers on X for each of the three categories.

Collecting posts from the selected accounts. We use the X Pro API's full-archive search endpoint to collect posts created by the selected accounts between eight weeks before and four weeks after likes were made private. As a result, we successfully collect 154,122 posts across 1068 accounts (see details in Table 5.1).

Difference-in-Differences Estimation

We employ *Difference-in-Differences (DiD)* to estimate the effect of private likes on users' liking behavior toward posts from high-reputational-risk accounts. We combine two complementary perspectives:

- **Between-group comparison.** We compare liking patterns between *high-reputational-risk* and *low-reputational-risk* accounts to test whether the visibility change disproportionately affected engagement with socially sensitive posts. We assume that the policy's impact should be larger for high-reputational-risk accounts, those for which public engagement carries greater reputational risks, making them our treatment group. Low-reputational-risk accounts, for which engagement is less socially costly, serve as the control group. We then estimate the differential change in the number of likes these two groups received before and after the policy shift (see Fig. 5.1).
- **Within-group comparison across engagement types.** We further compare liking behavior with other forms of engagement that were not directly affected by the policy change. Likes are strongly correlated with reposts ($r = 0.853$, $p < 0.001$; 95% CI: [0.852, 0.854]). Restricting our analysis to high-reputational-risk accounts, we thus examine whether likes increased relative to reposts after the policy shift, isolating the effects specific to private visibility from broader engagement dynamics (see Fig. 5.1).

The *parallel trends* is a key assumption for DiD identification: in the absence of the policy intervention, the difference between the treatment and control outcomes would have evolved similarly over time. In our first specification, this implies that the gap in likes between high- and low-reputational-risk accounts would have remained stable absent the visibility change; in the second, that likes and reposts would have evolved in parallel for high-reputational-risk accounts. Together, these complementary designs allow us to attribute observed shifts in liking behavior to the introduction of private like visibility rather than to general engagement trends on the platform.

We employ a DiD approach with both weekly leads-and-lags and aggregated pre-post specifications. The leads-and-lags model estimates the weekly dynamics of engagement before and after the policy change, allowing us to assess the parallel trends assumption and track the temporal evolution of treatment effects. The aggregated pre-post model captures the overall effect of making likes private on $\mathbb{X}$. High-reputational-risk posts or likes serve as the treatment group, with low-reputational-risk posts or reposts acting as controls. Additionally, post-specific and week-specific fixed effects are included in both models to account for unobserved heterogeneity across posts and time. Given that likes and reposts are count variables, we estimate our models using negative binomial regression. To interpret the results, we exponentiate the DiD estimates to obtain the relative change in likes attributable to the policy change. Further model details, including the full leads-and-lags and aggregated specifications, are in Suppl. S5.2.

5.3.2 Empirical Results

We start at examining the weekly averages of likes on high-reputational-risk posts relative to (i) the weekly averages of likes on low-reputational-risk posts (Fig. 5.2a) and (ii) the weekly averages of reposts on high-reputational-risk posts (Fig. 5.2b). The observation window spans from -8 weeks before to $+4$ weeks after the policy change. Across both comparison, we observe that weekly engagement levels remain relatively stable following the shift to private like visibility, with no clear divergence between high-reputational-risk vs. low-reputational-risk groups or likes vs. reposts. These descriptive patterns provide an initial indication that

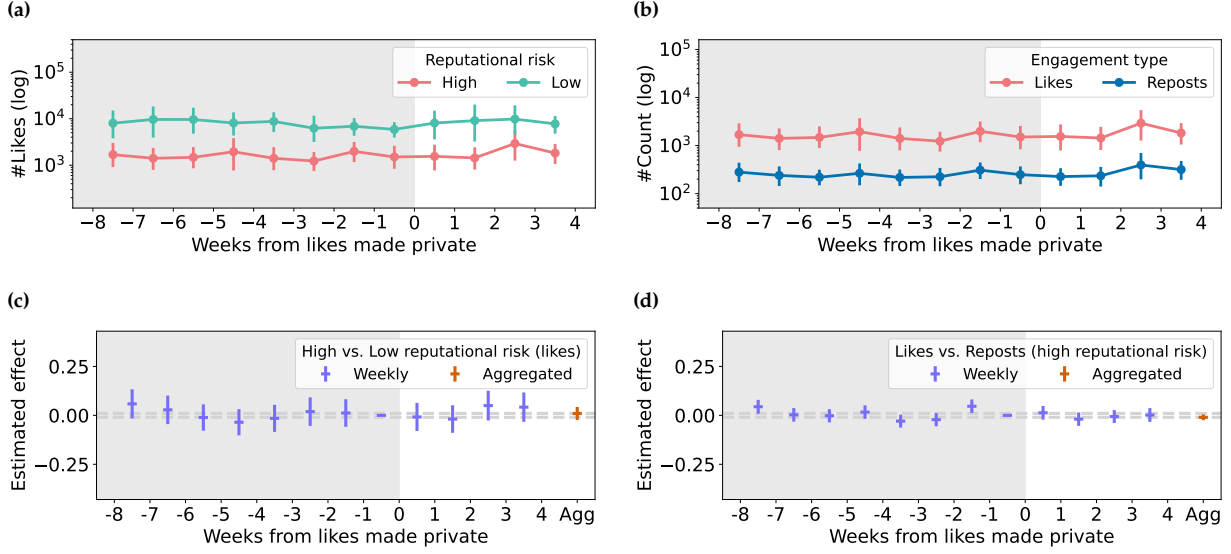


Figure 5.2: Analysis of weekly effects of making likes private on the number of likes. (a) The weekly averages of likes received by high-reputational-risk and low-reputational-risk posts. (b) The weekly averages of likes and reposts for high-reputational-risk posts. (c) The estimated weekly (and aggregated) effects of private visibility on likes to high-reputational-risk posts relative to low-reputational-risk posts. (d) The estimated weekly (and aggregated) effects of private like visibility on the number of likes for high-reputational-risk posts relative to the number of reposts. The error bars represent 95% CIs. The grey dash lines in (c) and (d) represent equivalence bounds.

making likes private did not trigger a large-scale shift in how users engage with high-reputational-risk content.

Weekly and aggregated DiD estimates. Fig. 5.2c reports the estimated weekly and aggregated effects of private visibility on the number of likes received by high-reputational-risk posts relative to low-reputational-risk posts. We find that all weekly effect estimates are close to zero (each $p > 0.5$), indicating no statistically significant change in likes to high-reputational-risk content following the policy change. The aggregated DiD estimate ($effect = 0.010$, $p = 0.564$; 95% CI: $[-0.023, 0.043]$) likewise shows a statistically non-significant effect and an effect size close to zero, implying that the overall effect of making likes private is negligible in both statistical and practical terms.

Similarly, we estimate the weekly and aggregated effects of private visibility on the number of likes relative to reposts for high-reputational-risk posts (Fig. 5.2d). We find that the weekly effect estimates are not significantly different from zero, although we observe minor pre-trends exist in week -8 ($effect = 0.044$, $p = 0.013$; 95% CI: $[0.009, 0.080]$) and week -2 ($effect = 0.046$, $p = 0.008$; 95% CI: $[0.012, 0.082]$). The aggregated effect estimate ($effect = -0.010$, $p = 0.180$; 95% CI: $[-0.025, 0.005]$) is likewise small and not statistically significant, consistent with the null results in the comparison of likes between high-reputational-risk and low-reputational-risk posts.

Equivalence test and HonestDiD estimates. To assess the robustness of our weekly DiD estimates, we (i) conduct an equivalence test to determine whether the non-statistically significant effects can be dismissed as practically null, and (ii) account for potential violations of the parallel trends assumption by applying the HonestDiD framework to obtain robust confidence intervals for the original estimates [77, 98]. For the equivalence test, we set an effect-size bound interval of $[-0.01, 0.01]$, which reflects the smallest change in weekly liking behavior that we

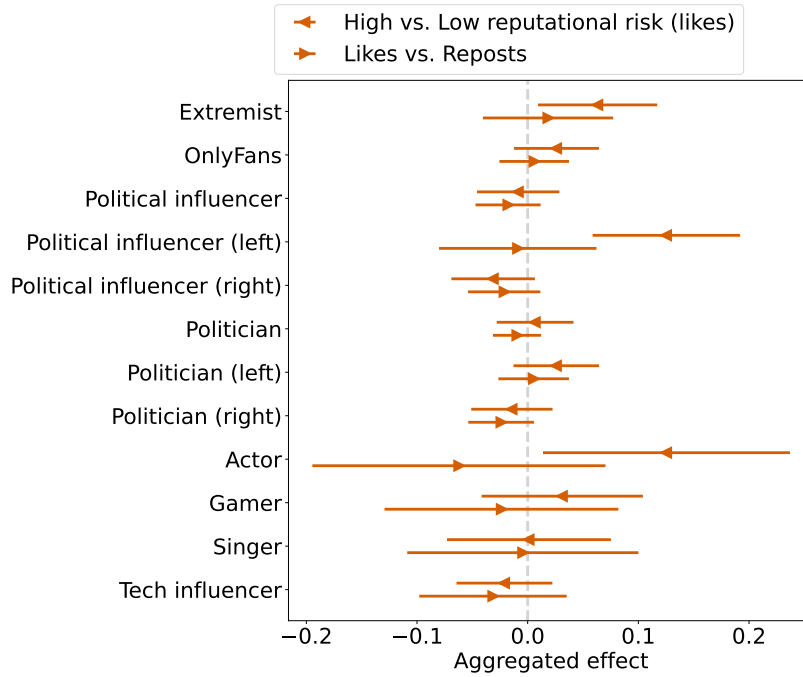


Figure 5.3: The estimated aggregated effects across account categories in Study 1. The error bars represent 95% CIs.

consider meaningful in this context. All weekly estimates fall within the bound interval (each $p > 0.05$; see Fig. 5.2c), implying a true null effect.

Additionally, HonestDiD provides a framework for robust inference in DiD settings where the strict parallel trends assumption may not fully hold [76, 241]. Rather than invalidating results when minor pre-trends are detected, HonestDiD allows for small deviations between treatment and control groups and adjusts the inference accordingly. Specifically, we implement the Conditional Least Favorable Hybrid (C-LF) method, assuming that any deviations from parallel trends in the post-treatment weeks do not exceed $\bar{M}$ times the largest deviation observed in the pre-treatment weeks. The results indicate that the robust estimates remain statistically indistinguishable from zero, reinforcing our finding that the private visibility did not significantly affect users' liking behavior toward high-reputational-risk content (see Fig. S6.5 in Suppl. S5.3 for details).

Aggregated effects across account categories and engagement types.

We conduct sensitivity analysis to examine how the aggregated effects of private visibility on the number of likes vary across account categories and engagement types (see Fig. 5.3). When comparing likes to high-reputational-risk posts vs. low-reputational-risk posts, we find that the estimated aggregated effects are significantly positive for SPLC extremists ($effect = 0.063$, $p = 0.018$; 95% CI: [0.011, 0.118]), left-leaning political influencers ($effect = 0.125$, $p < 0.001$; 95% CI: [0.061, 0.194]), and actors ($effect = 0.125$, $p = 0.019$; 95% CI: [0.019, 0.243]), while other aggregated DiD estimates are not statistically significant (each $p > 0.05$). However, importantly, when comparing likes with reposts for the same high-reputational-risk posts, the estimated aggregated effects are consistently statistically indistinguishable from zero (each $p > 0.05$). This suggests that the observed moderate effects in the high-reputational-risk vs. low-reputational-risk comparison for likes do not reflect a genuine response to the private visibility of likes, but instead stem from broader category-level differences unrelated to the policy change.

5.4 Study 2: User Study

To validate our findings observed in Study 1, we further conduct a user study to investigate whether the awareness of private visibility policy influences participants' liking behavior.

5.4.1 Experiment Setup

We design a within-subject survey experiment in which each participant was shown with topics that varied in *reputational risk*—that is, the potential reputational cost of publicly engaging with them. For each post, participants rated how likely they would be to “like” it under two conditions: (i) when likes were *publicly visible* and (ii) when likes were *private* (visible only to the author and the liker). This design allows us to compare self-reported willingness to engage across visibility settings and exposure levels within the same individuals.

The set of topics is selected to capture both high- and low-reputational-risk contexts. Controversial, high-reputational-risk topics include *Abortion*, the *Black Lives Matter (BLM) movement*, *Immigration and border security*, *Transgender rights*, and *Gun control*. Additionally, non-controversial, low-reputational-risk topics include *Art*, *Charities working to end hunger*, and *Pop culture*.

We recruit $N = 203$ participants through the Prolific platform, restricting the sample to U.S.-based users who reported being active $\mathbb{X}$ /Twitter users. To approximate the demographic composition of the U.S. $\mathbb{X}$ user base, we employ a stratified sampling procedure based on *age*, *household income*, and *sex*, using population estimates reported in the 2025 Pew Research Center survey on social media use [242]. While exact stratification is not possible due to platform-level constraints, we subsequently apply iterative proportional fitting (raking) to reweight the sample to match Pew's demographic margins (see Suppl. S5.4).

All procedures were reviewed and approved by the authors' institutional ethics board.

5.4.2 Experimental Results

Fig. 5.4 illustrates how participants first became aware of the private visibility policy on the liking behavior. We find that approximately half of participants ($mean = 0.434$; 95% CI: [0.356, 0.518]) reported learning about the change through news coverage or posts on $\mathbb{X}$, while a comparable portion ($mean = 0.465$; 95% CI: [0.383, 0.546]) indicated that they first became aware of the policy change through our survey. This finding suggests that although the policy change received moderate public attention when the visibility became private, a substantial portion of users may not have been immediately informed or conscious of the modification in the visibility of their likes. In the following analysis, we analyze participants' self-rated likelihood of liking posts on both high-reputational-risk and low-reputational-risk topics. Additionally, we investigate whether participants who were previously aware of the policy change through news/posts (*i.e.*, media-informed participants) and those who learned about it through other channels, *e.g.*, our survey, behave differently to compare the long-term and short-term effects of the private visibility.

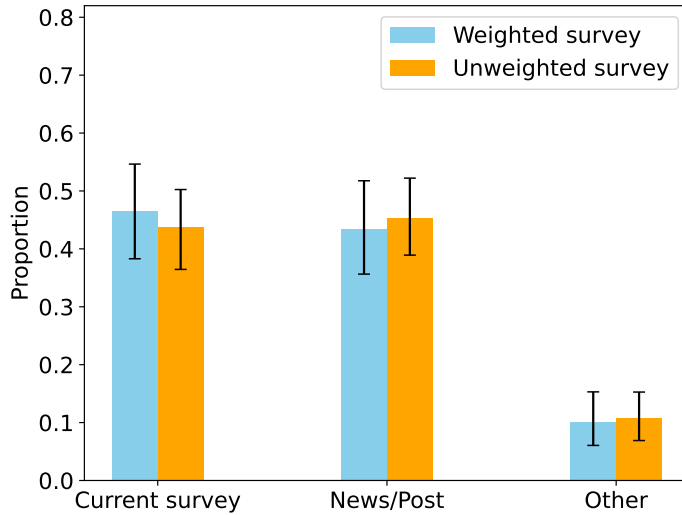


Figure 5.4: Reported channels through which participants first became aware of X's change to private like visibility. Error bars represent 95% CIs.

The likelihood of liking posts. The average participants' self-rated likelihood of liking posts (*i.e.*, group means) across high-reputational-risk and low-reputational-risk topics under public and private visibility conditions are presented in Fig. 5.5a. For all high-reputational-risk topics, we find that the average self-rated willingness to like related posts is slightly higher under private visibility than under public visibility. For instance, willingness increases (i) from 35.481 ($p < 0.001$; 95% CI: [30.602, 40.360]) to 40.260 ($p < 0.001$; 95% CI: [35.208, 45.313]) for posts related to abortion; (ii) from 34.382 ($p < 0.001$; 95% CI: [29.253, 39.510]) to 37.663 ($p < 0.001$; 95% CI: [32.513, 42.813]) for posts related to BLM movement; (iii) from 38.567 ($p < 0.001$; 95% CI: [33.772, 43.361]) to 43.699 ($p < 0.001$; 95% CI: [38.924, 48.474]) for posts related to immigration & border security; (iv) from 46.971 ($p < 0.001$; 95% CI: [42.061, 51.881]) to 51.275 ($p < 0.001$; 95% CI: [46.290, 56.261]) for posts related to corruption in the government; (v) from 29.473 ($p < 0.001$; 95% CI: [24.526, 34.420]) to 45.491 ($p < 0.001$; 95% CI: [40.569, 50.412]) for posts related to transgender rights; and (vi) from 41.942 ($p < 0.001$; 95% CI: [36.918, 46.966]) to 45.491 ($p < 0.001$; 95% CI: [40.569, 50.412]) for posts related to gun control. However, the confidence intervals for the two conditions significantly overlap across these topics, indicating that making liker identities anonymous does not substantially alter participants' self-reported willingness to engage with controversial content.

For low-reputational-risk topics, we observe an even larger overlap between the public and private conditions than for high-reputational-risk topics, suggesting that making likes private has minimal impact on engagement with non-controversial content. For example, the mean likelihood of liking posts related to art is 57.572 ($p < 0.001$; 95% CI: [53.080, 62.064]) in the public mode and 58.738 ($p < 0.001$; 95% CI: [54.215, 63.261]) in the private mode. For charities working to end hunger, the means are 54.471 ($p < 0.001$; 95% CI: [49.566, 59.375]) in the public mode and 54.081 ($p < 0.001$; 95% CI: [49.161, 59.000]) in the private mode. Additionally, for pop culture, we observe a pattern similar to that we see in high-reputational-risk topics: participants' self-rated willingness slightly increases from 49.586 ($p < 0.001$; 95% CI: [45.031, 54.141]) to 53.248 ($p < 0.001$; 95% CI: [48.679, 57.818]), while this increase remains limited and does not indicate a substantial effect of private likes.

Notably, we find that the average participants' self-rated likelihood of liking posts across all topics center around 50 out of 100, suggesting

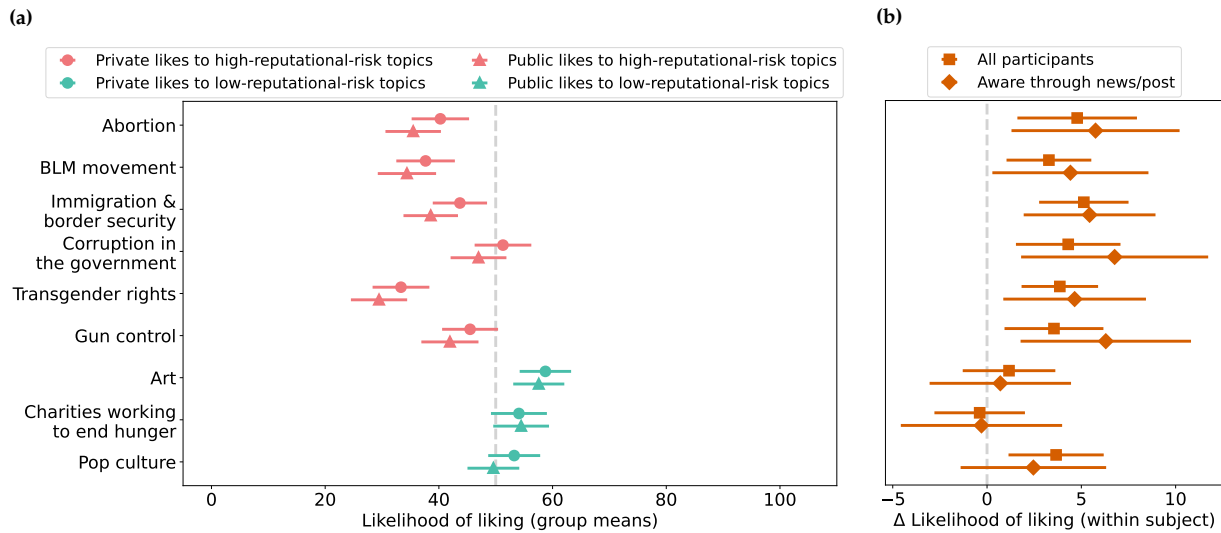


Figure 5.5: Participants' self-rated likelihood of liking posts on high-reputational-risk and low-reputational-risk topics if their likes were public or private ($N = 203$). (a) Mean likelihood of liking (0–100 scale) for each topic when likes were public or private (group means). (b) Mean within-subject change (Δ) in likelihood of liking between private and public conditions, plotted for all participants and for the subset who were aware of the platform's policy change through news or posts. Positive values indicate greater liking when likes were private. The error bars represent 95% CIs.

a generally moderate inclination to engage with content through likes. Meanwhile, as shown in Fig. 5.5a, the likelihood of liking posts related to high-reputational-risk topics remains consistently lower than that for low-reputational-risk topics, regardless of the visibility condition.

The within-subject change in likelihood of liking posts. At the group level, we find no evidence that private visibility significantly alters participants' self-reported willingness to engage with high-reputational-risk content. To capture more nuanced behavioral variation, we further examine the within-subject changes in participants' self-rated likelihood of liking posts between the public and private visibility conditions. This analysis allows us to assess whether individual participants, rather than aggregate group averages, exhibit shifts in engagement tendencies following the visibility change.

Specifically, we examine the within-subject changes by calculating the paired differences in each participant's likelihood of liking posts in the private condition relative to the public condition across high- and low-reputational-risk topics. As shown in Fig. 5.5b, we find that, for high-reputational-risk topics, the within-subject changes in participants' self-rated likelihood of liking posts consistently show small but statistically significant increases from public condition to private condition across abortion ($mean = 4.780$, $p = 0.003$; 95% CI: [1.609, 7.950]), BLM movement ($mean = 3.281$, $p = 0.004$; 95% CI: [1.034, 5.529]), immigration & border security ($mean = 5.133$, $p < 0.001$; 95% CI: [2.762, 7.503]), corruption in the government ($mean = 4.304$, $p = 0.002$; 95% CI: [1.533, 7.076]), transgender rights ($mean = 3.860$, $p < 0.001$; 95% CI: [1.830, 5.889]), and gun control ($mean = 3.549$, $p = 0.008$; 95% CI: [0.926, 6.172]). Across low-reputational-risk topics, the within-subject changes for art ($mean = 1.166$, $p = 0.350$; 95% CI: [−1.289, 3.622]) and for charities working to end hunger ($mean = -0.390$, $p = 0.749$; 95% CI: [−2.790, 2.010]) are not statistically significant. However, for pop culture, the within-subject change is significantly positive ($mean = 3.663$, $p = 0.005$; 95% CI: [1.138, 6.187]). This suggests, for all participants, the within-subject changes

are generally negligible across low-reputational-risk topics, with the exception of pop culture, which may also trigger mild self-censorship.

Additionally, we conduct a subset analysis for media-informed participants who learned about the policy change through news coverage or posts on $\mathbb{X}$. Given that the publication of news items or posts related to policy are time-sensitive, these media-informed participants were likely active online and aware of the visibility change shortly after its implementation. Consequently, focusing on media-informed participants can help us examine a relatively long-term effect compared to those who only learned about the policy change during the survey. Fig. 5.5b shows that the within-subject changes among media-informed participants are significantly positive and consistent with those observed across all participants for high-reputational-risk topics: abortion ($mean = 5.753, p = 0.012$; 95% CI: [1.296, 10.210]), BLM movement ($mean = 4.420, p = 0.037$; 95% CI: [0.283, 8.558]), immigration & border security ($mean = 5.438, p = 0.003$; 95% CI: [1.943, 8.933]), corruption in the government ($mean = 6.766, p = 0.008$; 95% CI: [1.807, 11.726]), transgender rights ($mean = 4.645, p = 0.017$; 95% CI: [0.862, 8.427]), and gun control ($mean = 6.297, p = 0.007$; 95% CI: [1.778, 10.816]). On the contrary, the within-subject changes among media-informed participants are statistically not significant across low-reputational-risk topics: art ($mean = 0.704, p = 0.710$; 95% CI: [-3.046, 4.454]), charities working to end hunger ($mean = -0.295, p = 0.891$; 95% CI: [-4.572, 3.982]), and pop culture ($mean = 2.457, p = 0.209$; 95% CI: [-1.401, 6.316]).

In summary, the within-subject changes consistently show that the participants' self-rated likelihood of liking posts significantly increases from the public to the private condition across all high-reputational-risk topics, while the within-subject changes are not statistically significant across non-controversial topics, particularly among media-informed participants. These findings suggest that making likes private indeed increases participants' willingness to engage with high-reputational-risk posts. However, these within-subject increases are not substantial enough to result in a meaningful change in the overall likelihood of liking posts, thereby not observable at the group level.

5.5 Discussion

Social media platforms provide important venues for users to interact with others and manage their own personas. Yet, these behaviors are shaped and constrained by the design and visibility of engagement features on the platforms [195, 198]. Recently, $\mathbb{X}$ implemented a major design change on its engagement feature, making previously public likes private, which means that likers are only visible to post authors and themselves. This change raises an important question: *Does the private visibility encourage users to like sensitive or high-reputational-risk content more often than the public visibility?* Here, we conduct a comprehensive analysis from both the platform side and the user side. We find that the private visibility did not imply a substantial change in the number of likes for high-reputational-risk posts on $\mathbb{X}$. While we observe that users' self-reported willingness in liking controversial posts increased under the private condition, this increase did not result in a meaningful effect in behavior change.

Given that our analysis considers changes in the number of likes on high-reputational-risk posts within the four weeks following the imple-

mentation of private visibility, it is possible that some users were not yet aware of this change during this period. To address this concern, we assume that relevant posts and news coverage were published shortly after the visibility update and examine the patterns of participants who learned about the change through these sources, *i.e.*, media-informed participants. Our findings show that media-informed participants exhibited similar changes in willingness to like posts as other participants. Therefore, the lack of substantial changes in the number of likes to high-reputational-risk posts at the platform level is unlikely to be solely due to users' unawareness of the visibility update. Moreover, the consistent patterns between media-informed participants and participants who were aware of the visibility change through our survey suggest that the willingness changes are stable rather than an immediate reaction to awareness.

Intention–behavior gap. The divergence between users' self-reported willingness and observed behavior suggests an *intention–behavior gap* in digital contexts. Previous research has shown that users are more willing to share controversial content under anonymous or private conditions [229], and that liking intentions have positive impacts on liking behavior [243, 244]. However, changing intentions does not guarantee behavior change, and there is a large intention-behavior gap: a medium-to-large-sized change in intentions lead to only a small-to-medium-sized change in behavior [245, 246]. Further evidence suggests that intentions get translated into action approximately one-half of the time [247]. Consistent to this intention–behavior gap, the modest increase in willingness to like posts under private visibility on X does not translate into substantial changes in actual engagement. Similar patterns have been observed in prior studies of platform design changes. For instance, in a case study of Twitter's character limit, Gligoric et al. [248] found that although post-intervention "crammed" tweets were longer, their syntactic and semantic characteristics remained largely unchanged, reflecting only superficial adjustments rather than meaningful behavioral shifts. Likewise, X temporarily suspended the retweet function and prompted users to use the quote tweet function during 2020 U.S. presidential election. However, the suspension of the retweet did not encourage users to post more quotation tweets or longer comments [249].

The concentration of engagement among heavy or automated accounts. We note that engagement is concentrated on social media platforms—a small fraction of highly active users and/or automated agent accounts for a large share of engagement [136, 202, 203, 250, 251]. This concentration can obscure individual-level behavioral trends, making subtle changes in engagement among ordinary users difficult to detect. In addition, engagement patterns are not only unequal but also strategically distributed. Users could maintain multiple or pseudonymous accounts that serve different social purposes, for instance, separating public-facing profiles from private or anonymous accounts [229, 252]. Users may use their secondary accounts to engage with content they perceive as controversial or reputationally risky, thereby decoupling their expressive behavior from their main identity. This behavioral adaptation can reduce the impact of visibility interventions.

Practical implications. Our findings suggest that visibility interventions aimed solely at reducing reputational risk may be insufficient to meaningfully shift platform-wide behavior. Instead, such changes may need to be paired with complementary design levers—such as algorithmic adjustments to content exposure, temporal interventions in feedback timing, and audience segmentation tools—to influence engagement in

more consequential ways. By *algorithmic adjustments*, platforms can decouple private engagements from recommendation logic, reducing the fear of algorithmic surveillance. Similarly, to further reduce users' fear of surveillance, platforms could use *temporal interventions* to alter when users receive engagement feedback, such as delaying or batching notifications about likes. By loosening the immediacy of feedback, these interventions weaken the sense that one's actions are being instantly observed and evaluated, thereby reducing the real-time reputational pressures associated with engaging with controversial content. Finally, using *audience segmentation tools*, platforms could allow users to control which followers or sub-audiences can see their interactions, enabling more context-specific engagement without fully public exposure.

Limitations and future research. As with all research, ours is not free of limitations and offers opportunities for future research. First, we focus on aggregate engagement metrics (likes and reposts) and do not directly observe the individual-level interactions, which may limit understanding of nuanced behavioral responses. Future work with access to such data could examine how users perceive and respond to hidden liker identities, and how their behavior unfolds over time. Second, engagement on X is highly concentrated among a small fraction of highly active or automated accounts, which may obscure subtle effects among ordinary users and limit the detectable power of platform-level analyses. Future work could attempt to distinguish ordinary users from heavy users and bots to assess how making likes private affects different segments of the user base. Third, our survey experiment relies on self-reported willingness to like posts, which may be subject to social desirability or hypothetical bias, and may not perfectly map onto actual behavior. Follow-up research could combine surveys with field interventions to more directly characterize the gap between what users say they would do under public vs. private like visibility and how they actually behave. Finally, our study focuses on a single platform (X) and specific content types (*e.g.*, high-reputational-risk and controversial posts). More research is needed to test the generalizability of our findings across other platforms, culture contexts, and engagement features.

Abstract: Community-based fact-checking is a promising approach to correct misleading posts at scale. Yet, causal evidence regarding its effectiveness in reducing the spread of misinformation on social media is missing. Here, we performed a large-scale empirical study to analyze whether community notes reduce the spread of misleading posts on X. Using a Difference-in-Differences design and repost time series data for $N = 237,180$ (community fact-checked) cascades that had been reposted more than 431 million times, we found that exposing users to community notes reduced the subsequent spread of misleading posts by, on average, 61.2%. The effect was pronounced across the board but significantly weaker for posts from influential accounts and political content. Additionally, community notes increased the odds that users delete their misleading posts by 94.3%. Although community notes were broadly effective in reducing the spread of posts once annotated, they often appeared too late to intervene in the early (and most viral) stage of the diffusion. As a result, their system-wide effect was more modest, lowering total engagement with misleading posts by 14.9%. Our work offers important implications to enhance the effectiveness of community-based fact-checking approaches on social media.

6.1 Introduction 105

6.2 Results 107

6.2.1 Difference-in-Differences Estimation 107

6.2.2 Effect on the Cumulative Repost Count 109

6.2.3 Analysis of Reposting Mechanisms 111

6.2.4 Effect on Post Deletion and Suspension/Protection 112

6.3 Discussion 114

6.4 Methods 116

6.4.1 Data Collection 116

6.4.2 Empirical Analysis 116

Publication
Y. Chuai et al., ‘Community-based fact-checking reduces the spread of misleading posts on X (formerly Twitter),’ <i>Nature Communications</i> , 2026, forthcoming.

6.1 Introduction

The spread of misinformation on social media platforms has become a concerning issue of the digital age and raised alarm bells across various domains. Negative repercussions of misinformation have been repeatedly observed, with tangible consequences in critical areas such as elections [21, 250, 253–256], public health [25, 257–259], and public safety [141, 260, 261]. Researchers, governments, and regulation authorities thus urge social media providers (e.g., X/Twitter, Facebook) to develop effective countermeasures to counteract the spread of misinformation on their platforms [5, 155, 262–264].

To this end, a widely implemented approach is the use of professional fact-checkers to identify and label misleading posts [265, 266]. The rationale is that if users are warned that a message is false, they should be less likely to believe it. While this approach has been shown to be effective in numerous experimental studies [44, 170, 190, 267–274] (for a review, see [44]), it faces several key challenges, particularly with regard to volume, visibility, and trust. First, in-depth investigation of claims can be time-consuming, often taking many hours or even days [275]. Given the limited number of available fact-checkers, not every claim can be thoroughly examined. Consequently, fact-checkers are compelled to prioritize content that is overtly false or intentionally misleading over material that is more intricate or nuanced [53]. Second, professional fact-checks frequently have

a restricted audience. Except for some collaborations with social media platforms on particular topics, fact-checking organizations primarily disseminate their findings through their own websites. A study conducted in 2017 revealed that over half of all U.S. adults had never visited any fact-checking website [276], demonstrating the limited reach of professional fact-checks. Third, a major drawback is that professional fact-checkers face a lack of trust from large portions of society [277, 278]. For example, according to surveys, the majority of Republican partisans (70%) and half of all U.S. adults believe that fact-checkers exhibit bias [278].

As a remedy, research has proposed to employ non-expert fact-checkers in the crowd to fact-check social media content [46, 47, 51–53, 279–281]. This approach is based on the idea that the biases and errors of a single user can be mitigated by tapping into the collective intelligence of diverse individuals, also known as the “wisdom of the crowds” [281–283]. The wisdom of crowds has been repeatedly observed in a wide range of settings, including online platforms such as Wikipedia and Stack Overflow, where the crowd ensures relatively trustworthy and high-quality accumulation of knowledge [284]. Experimental studies on crowd-based fact-checking found that the assessments of even relatively small crowds are comparable to those of experts [46, 47, 51–53, 281, 285]. Furthermore, crowd-based fact-checking enables the examination of significantly larger volumes and varieties of posts as compared to expert-based assessments [96, 150, 167, 286]. Altogether, community-based fact-checking may have the potential to address many of the drawbacks of expert-based fact-checking on social media.

Building upon these encouraging findings, the social media platform X (formerly Twitter) has recently introduced “Community Notes” (formerly “Birdwatch”), a community-based fact-checking system that allows users to assess the accuracy of posts [167, 287]. “Community Notes” is the first large-scale attempt to implement community-based fact-checking on a major social media platform. It enables enrolled users to flag posts that they believe are misleading and add concise fact-checking explanations, which are then rated by other users for their helpfulness [288]. Notes rated by the community to be helpful then appear directly on the fact-checked misleading post to inform other users. However, empirical evidence on whether displaying community notes reduces the spread of misleading posts on social media is missing so far.

In prior research, the efficacy of fact-checking has mainly been studied in one of two forms: (i) lab/survey experiments evaluating whether flagging misinformation influences misinformation discernment and sharing intentions [56, 170, 190, 268, 269, 271–273], and (ii) field studies analyzing the spread of misleading vs. non-misleading posts [14, 74, 96, 289, 290]. The former group of studies, *i.e.*, lab/survey experiments, have shown that misinformation flags are effective in increasing the participants’ ability to discern between true and false content [56, 190, 271, 272] as well as decreasing their sharing intentions for misinformation [56, 268, 269, 273]. The latter of studies, *i.e.*, field studies, were largely limited to correlational works analyzing the virality of misleading vs. non-misleading posts on social media [14, 34, 74, 81, 96, 167, 289, 290]. As an example, Vosoughi et al. [14] found that expert fact-checked misinformation is linked to more viral resharing cascades. Still, these works have not quantified the real-world intervention effect of fact-checking, mainly due to data constraints such as missing time series data on reposts and interventions. Here, we add by analyzing the efficacy of community notes in reducing the spread of misleading posts on the social media platform X.

In this work, we performed a large-scale quasi-experimental study to analyze whether community fact-checks reduce the spread of misleading posts on the social media platform $\mathbb{X}$. For this, we examined $N = 237,180$ (community fact-checked) $\mathbb{X}$ cascades that were created within a period of over 20 months from the roll-out of “Community Notes” on October 6, 2022 to June 11, 2024, and had been reposted more than 431 million times. To evaluate the efficacy of community notes in reducing the spread of misleading posts on $\mathbb{X}$, we collected *time series data* of repost counts for the community fact-checked posts over 36 hours since their creation. We then employed a Difference-in-Differences (DiD) design and negative binomial regression models to estimate the reduction in reposts for misleading posts after the display of community notes. Overall, this setting offers a unique opportunity to study the real-world efficacy of community-based fact-checking on a major social media platform and enables us to quantify the intervention effect of community notes.

As a secondary analysis, we examined the extent to which the “Community Notes” feature in its current implementation reduced the spread of misleading posts on $\mathbb{X}$ in terms of cumulative repost count. The underlying rationale is that the display timing of the notes might be critical. Even if users are less inclined to share tweets once they are equipped with a community note, the fact-check might arrive too late to intervene in the early and most viral stage of the diffusion, thereby potentially diminishing its effect on overall engagement with misinformation on social media.

6.2 Results

6.2.1 Difference-in-Differences Estimation

We employed a Difference-in-Differences (DiD) design and negative binomial regression models to estimate the Average Treatment effect on the Treated (ATT), *i.e.*, the extra repost reduction in the treatment group after the display of community notes compared to the baseline before the display and relative to the control group (see Section 6.4). To mitigate potential confounding factors and balance treatment and control groups, we constructed a control group from the source posts without displayed notes through one-to-one matching. The one-to-one matching was conducted based on the variables from user profiles (*e.g.*, followers and followees) and post features (*e.g.*, sentiments and topics). As a result, the source posts in the control group have no statistically significant difference with the source posts in the treatment group with respect to the user and post characteristics (see details in Suppl. S6.4). Subsequently, we assigned the virtual time of note display to the source posts in the control group according to the corresponding posts in the treatment and recenter the repost timelines of the source posts around the display of community notes in the treatment and control groups.

We started our analysis with a two-period DiD model. This allowed us to estimate difference in outcomes (*i.e.*, repost counts) between the treatment and control group before and after the treatment (*i.e.*, displaying community notes) was applied. To this end, we considered the period between 1 and 12 hours from the note display as the after-display period and the period between 1 to 4 hours before the note display as the before display period (see Section 6.4). The ATT for the two-period DiD model is visualized in purple color in Fig. 6.1 (see Suppl. S6.5 for full estimation

results). The estimate for the ATT was -0.612 (99% CI: $[-0.617, -0.608]$; $p < 0.001$), which indicates that displaying community notes reduced the subsequent number of reposts by, on average, 61.2%. This suggests a pronounced treatment effects of community notes in reducing the spread of misleading posts on $\mathbb{X}$.

Subsequently, we estimated a multi-period DiD to enable a temporal analysis of the treatment effect. Here, we again considered a 4-hour window before the note display as the before-display period and then examined the hourly multi-period ATTs from 1 to 12 hours after the note display (see Section 6.4). Equivalence testing [291–293] confirms that the ATT estimates in the hours before note display were small enough to be considered negligible, indicating the absence of meaningful pre-trends and supporting the validity of the DiD framework (see Suppl. S6.5.2). Fig. 6.1 visualizes the hourly multi-period ATTs over time (see Section S6.5 for full estimation results). We observed statistically significant and pronounced ATTs after the display of community notes. During the first one hour after the note display, the ATT was estimated as -0.363 (99% CI: $[-0.380, -0.346]$; $p < 0.001$). This implies that community notes reduced the number of reposts by 36.3% within the first hour. Furthermore, the efficacy of community notes increased over time following the note display. The ATTs were -0.530 (99% CI: $[-0.542, -0.517]$; $p < 0.001$) at the second hour, -0.613 (99% CI: $[-0.623, -0.602]$; $p < 0.001$) at the fourth hour, and -0.644 (99% CI: $[-0.654, -0.634]$; $p < 0.001$) at the eighth hour. At the twelfth hour after the note display, the hourly ATT was -0.674 (99% CI: $[-0.683, -0.664]$; $p < 0.001$). In sum, our temporal analysis indicates that community notes reduced the number of reposts from 36.3% to 67.4% within the first 12 hours after they were displayed to users.

We conducted a comprehensive set of checks to ensure the robustness of our DiD estimates: (i) We performed a placebo test (Suppl. S6.6) by replacing the treatment group with a placebo group. The results showed no additional reduction in reposts within the placebo group, which supports that the estimated ATTs in the treatment group can be attributed to the display of community notes. (ii) To assess the impact of potential violations of parallel trends, we applied HonestDiD, a robust inference approach that quantifies uncertainty in DiD estimates [292, 294]. Our estimates remained stable even when allowing post-display deviations to be twice as large as the largest pre-display deviation (Suppl. S6.5.2). This demonstrates that our findings hold even under relaxed parallel trends assumptions. (iii) We implemented a battery of additional checks, including alternative model specifications, estimators, and timeframes (Suppl. S6.9), all of which consistently supported the robustness of our findings.

Furthermore, we complemented our main analysis with a wide range of sensitivity analyses (Fig. 6.2). First, we studied how the efficacy of community notes varied depending on the post age at note display (Fig. 6.2a). Here, we found that the efficacy of community notes was higher if they were displayed earlier. Second, we examined changes in the efficacy of community notes since the roll-out of the feature in October 2022. As shown in Fig. 6.2b, we observed a trend of increasing efficacy of community notes in the months following its launch. This may be partly attributed to improvements within the Community Notes program, such as faster note generation (see Suppl. S6.11). Third, we examined the role of helpfulness ratings (Fig. 6.2c) and found that community notes with higher helpfulness scores (i.e., higher agreement among raters) had a higher efficacy. Fourth, we analyzed whether the treatment

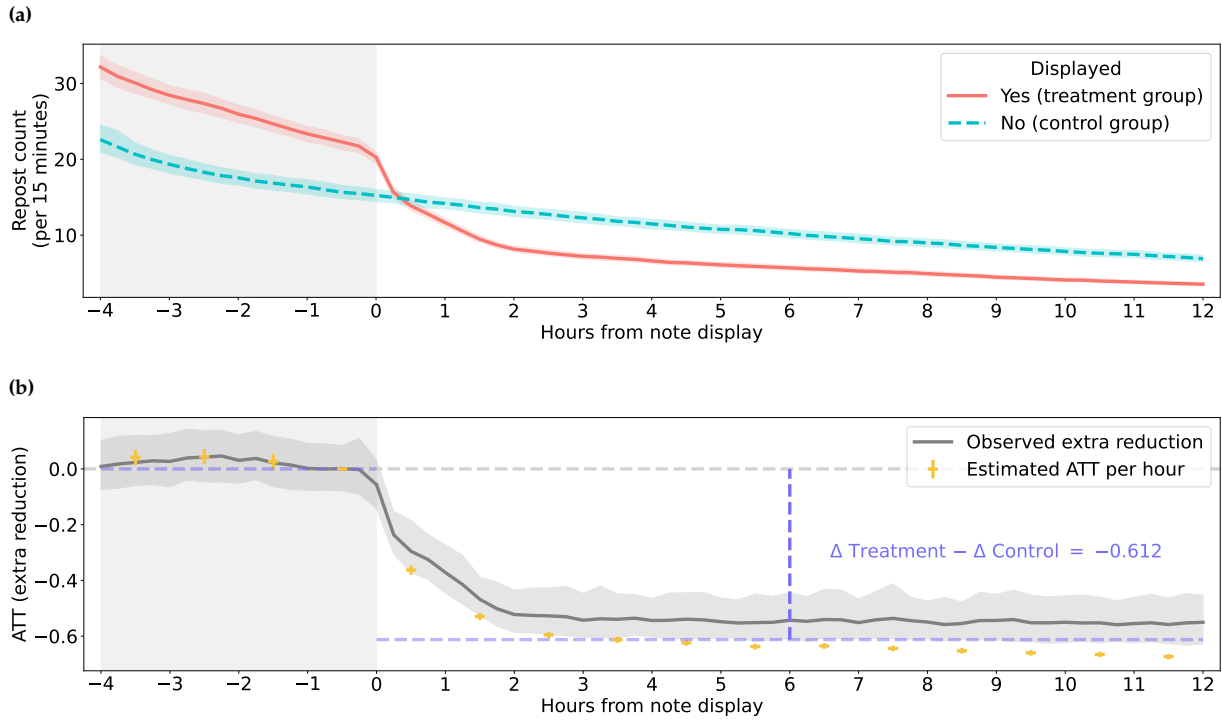


Figure 6.1: Community notes reduce the spread of misleading posts on X. (a) Time series of repost counts within 15 minutes intervals in the treatment group (red) and the control group (blue) from 4 hours before the display of community notes to 12 hours after the display of community notes. The error bands represent 99% Confidence Intervals (CIs). (b) Two-period (purple) and multi-period (yellow) ATTs estimated using a Difference-in-Differences (DiD) design and negative binomial regression models. For the two-period DiD model, the ATT is calculated as $\Delta \text{Treatment} - \Delta \text{Control} = -0.612$. For the multi-period DiD model, the yellow circles and error bars show the estimated hourly multi-period ATTs (with 99% CIs). The grey band (with 99% CIs) visualizes the observed extra reduction of the ratio of reposts in the treatment group relative to reposts in the control group and compared to the ratio of reposts before the display of community notes. The ATT estimations are based on 654,400 repost time series observations for $N = 40,900$ posts. Post-level random effects are included. Full estimation results are in Tables S6.4–S6.5.

effect of community notes was moderated by author characteristics (*e.g.*, the number of followers, verified status) and post characteristics (*e.g.*, sentiment, topics; see Fig. 6.2d). We found that the efficacy of community notes was lower for source posts from verified users and high-follower accounts. This suggests that the effectiveness of community notes might be slightly discounted for source posts from accounts with high social influence. Regarding post characteristics, we observed that community notes had a lower efficacy for posts that had media elements and if they were they were attached to posts covering health-related, and political topics. Notably, despite these statistically significant variations (each $p < 0.001$), the treatment effect remained statistically significant and pronounced across the board. This points towards a broad effectiveness of community-based fact-checking on social media. The full results of our sensitivity analyses can be found in Suppl. S6.7.

6.2.2 Effect on the Cumulative Repost Count

The results of our DiD analysis imply that displaying community notes can reduce the spread of misleading posts on X by, on average, 61.2% compared to the control group and relative to the before-display period. However, compared to the rapid dissemination of posts on X, fact-checking via community notes was relatively slow (see Suppl. S6.1). While 75.7% of all notes deemed helpful were displayed to general users within 36 hours after their creation, the time lag between post creation

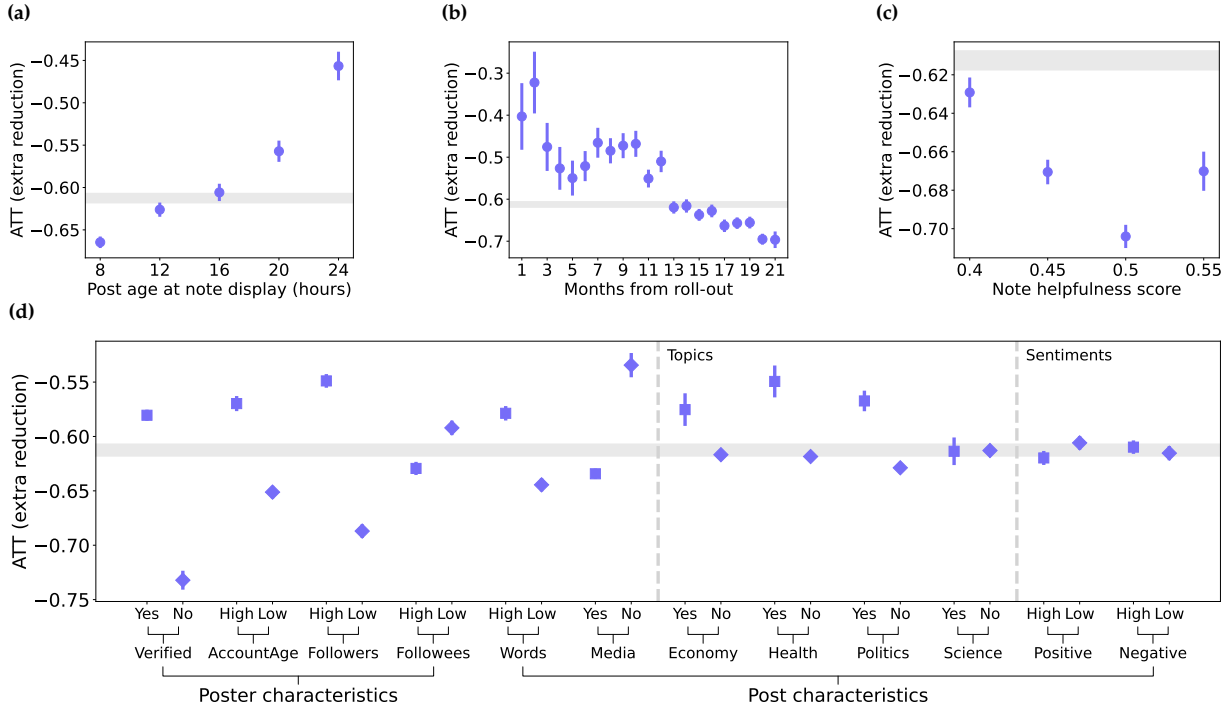


Figure 6.2: Sensitivity analysis. (a) The estimated ATTs across different response times from post creation to note display (grouped within 4-hour windows). Full regression results are reported in Table S6.10. (b) The estimated ATTs of community notes across the months following the roll-out of “Community Notes” program in October 2022. Full regression results are reported in Table S6.11. (c) The estimated ATTs of community notes depending on helpfulness scores. Full regression results are reported in Table S6.14. (d) The estimated ATTs within subgroups separated by poster and post characteristics. Full regression results are reported in Table S6.15, Table S6.16, and Table S6.17. In all plots, the error bars represent 99% CIs and the grey bands visualize the ATT (with 99% CIs) estimated via the two-period DiD model from our main analysis.

and note display (*i.e.*, the response time) averaged at 62.9 hours (median of 18.1 hours). In contrast, posts on $\mathbb{X}$ spread considerably faster: the average half-life (*i.e.*, the post age at which the cumulative ratio of reposts reaches to 50% of 36-hour reposts) amounted to merely 6.25 hours for posts with displayed notes. As a consequence, community notes, on average, might have come too late to reduce the spread of misleading posts at the most viral stages of diffusion. To further scrutinize the system-level effect of community notes, we analyzed the share of the cumulative repost count for misleading posts actually prevented by the community notes on $\mathbb{X}$.

To this end, we used our DiD models to predict the number of reposts (at 15-minute intervals) that the treated posts would have received in the absence of displayed notes. Subsequently, we compared the ratio of the predicted cumulative repost count (over 36 hours) in the absence of displayed notes to the actual cumulative number of reposts (over 36 hours). Fig. 6.3a shows the actual vs. predicted ratio across different response times from post creation to note display. As expected, we found that the ratio of reposts prevented due to community notes decreased over the response time, *i.e.*, was lower for late vs. early notes. The average reduction in the actual vs. predicted ratio is -0.149 , which means that community notes reduced the overall number of reposts on $\mathbb{X}$ by merely 14.9% (median of 10.1%). Similarly, in absolute numbers, we only observed a reduction of, on average, 430 reposts between the actual cumulative repost count (mean of 1792) compared to the predicted cumulative repost count (mean of 2222; $KS = 0.039$, $p < 0.001$; see Fig. 6.3b/c).

Additionally, we simulated the potential reduction in the cumulative

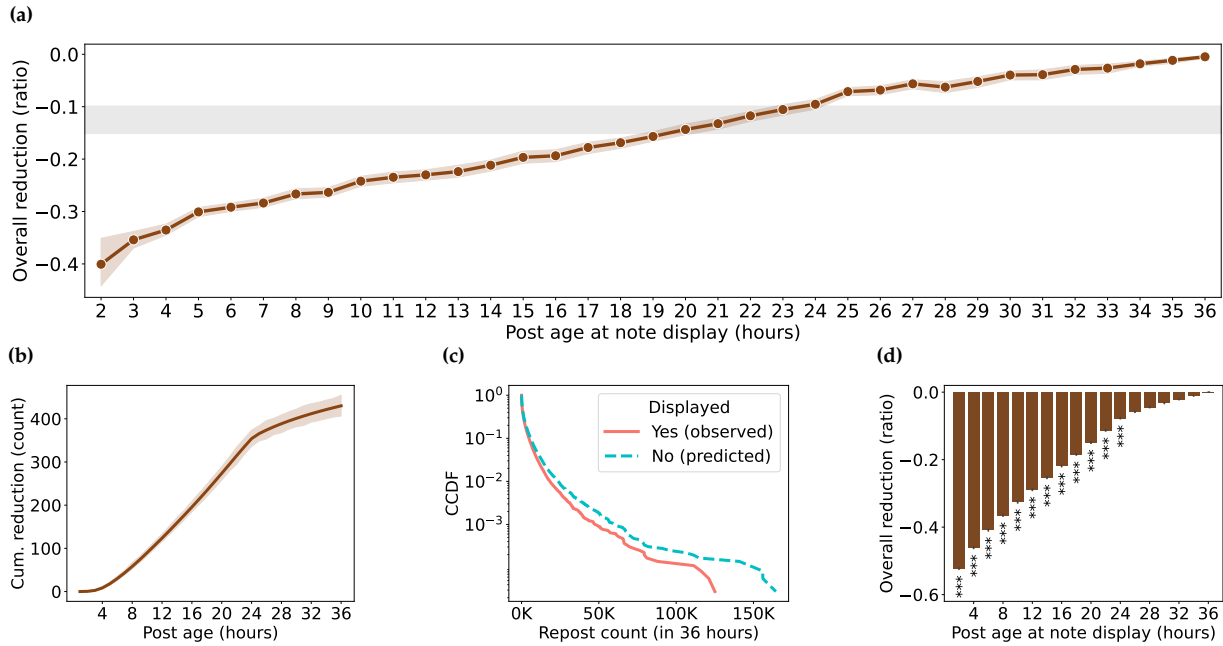


Figure 6.3: Effect of community notes on the cumulative repost count of misleading posts on X. (a) The changes in the ratio of the reduction of reposts relative to the predicted overall repost count over the response time from post creation to note display (grouped within 1-hour windows). The grey band ranges between the mean and the median of the ratio of the reduction. (b) The estimated cumulative count of reposts that community notes prevents at different post ages. The error band represents 99% CIs. (c) CCDFs showing the actually observed repost count for source posts with displayed community notes and the predicted repost count that the source posts would have received in the absence of community notes display. (d) The estimated overall reduction in reposts if all posts with community notes would have been displayed simultaneously from 2 to 36 hours after post creation. Statistical significance (* $p < 0.01$; ** $p < 0.005$; *** $p < 0.001$) was calculated using two-tailed KS tests.

repost count if all community notes *would* have been displayed at specific times from 2 to 36 hours after post creation. As shown in Fig. 6.3d, if community notes would have been displayed on misleading posts at the second hour since post creation, they would have reduced the overall repost count by 52.3% ($p < 0.001$). The reduction gradually decreased with increasing post ages at note display and became statistically insignificant if more than 24 hours had passed since the creation of the posts.

6.2.3 Analysis of Reposting Mechanisms

To shed further light on the causal mechanisms underlying the reduction in resharing of community noted posts, we used the X API to collect additional information on a random subset of users who reshared posts with community notes (see Suppl. S6.8). This reposter dataset (across $N = 3163$ posts) allowed us to analyze four potential mechanisms through which community notes may influence resharing behavior among reposters: (i) prior interactions with the authors of the fact-checked post (e.g., reduced sharing by dedicated followers), (ii) the verification status on X (e.g., reduced sharing by verified users), (iii) the political orientation of both posters and reposters (e.g., reduced sharing for politically concordant posts), and (iv) prior exposure to misinformation on X (e.g., reduced sharing by frequent consumers of misinformation).

The estimated ATTs across the four reposter characteristics are visualized in Fig. 6.4a (see Suppl. S6.8 for full estimation results). Our results imply that a key mechanism by which community notes reduced engagement was by curbing resharing by users with no prior interaction (*i.e.*, disconnected users) with the authors of the misleading posts. Specifically, the

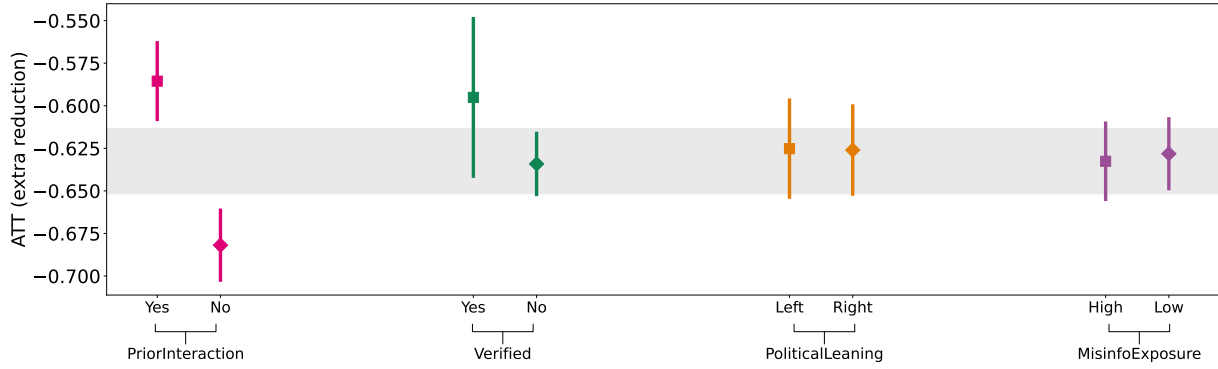


Figure 6.4: Analysis of reposting mechanisms. The estimated ATTs across different groups of reposters. Here we distinguish between (i) reposters that had prior interactions with the authors of the fact-checked post, (ii) the verification status of reposters, (iii) the political orientation of reposters, and (iv) prior exposure of reposters to misinformation on $\mathbb{X}$. The ATT estimations are based on a random subset of 50,608 repost time series observations for $N = 3136$ posts. Post-level random effects are included. Full estimation results are in Suppl. S6.8. In all plots, the error bars represent 99% CIs and the grey bands visualize the ATT (with 99% CIs) estimated via the two-period DiD model based on reposts from all categories of reposters.

efficacy of community notes was, on average, 9.6% higher for reposters who had no prior interactions with the authors of misleading posts (ATT: -0.682 ; 99% CI: $[-0.703, -0.660]$; $p < 0.001$), compared to reposters who had prior interactions (ATT: -0.586 ; 99% CI: $[-0.608, -0.561]$; $p < 0.001$). This suggests that community fact-checks may be less effective in tackling misinformation disseminated by influential accounts with dedicated and trusting audiences (*e.g.*, politicians, public figures).

In contrast, we did not find statistically significant differences in sharing behavior between verified (ATT: -0.595 ; 99% CI: $[-0.640, -0.545]$; $p < 0.001$) vs. non-verified reposters (ATT: -0.634 ; 99% CI: $[-0.653, -0.615]$; $p < 0.001$), left-leaning (ATT: -0.625 ; 99% CI: $[-0.654, -0.595]$; $p < 0.001$) vs. right-leaning reposters (ATT: -0.626 ; 99% CI: $[-0.652, -0.598]$; $p < 0.001$), and reposters with high (ATT: -0.633 ; 99% CI: $[-0.655, -0.609]$; $p < 0.001$) vs. low exposure to misinformation (ATT: -0.628 ; 99% CI: $[-0.649, -0.606]$; $p < 0.001$). This implies that displaying community fact-checks on misleading posts is broadly effective across both sides of the political spectrum and that even people who are steeped in misinformation are responsive to community fact-checks.

6.2.4 Effect on Post Deletion and Suspension/Protection

We further examined the effect of community notes on the likelihood of deletion of misleading posts by their authors. Notably, the $\mathbb{X}$ API only provided us with the deletion status and did not offer access to information regarding the time of deletion. Therefore, we did not know whether the posts were deleted before or after the display of community notes. Given this, we employed Regression Discontinuity Design (RDD) and hypothesized that, in the absence of a significant effect of community notes, the probability of post deletion should have exhibited smooth continuity (or keep stable) around the threshold of the note helpfulness score that determined the display of community notes (see Suppl. S6.10 for details). As of the date of data collection, the cut-off point (threshold) for the note helpfulness score was 0.40, meaning that only community notes receiving a score of 0.40 or above are eligible to be displayed on the corresponding misleading posts. Thus, in our RDD model, observations are assigned to the treatment or control group based on their values

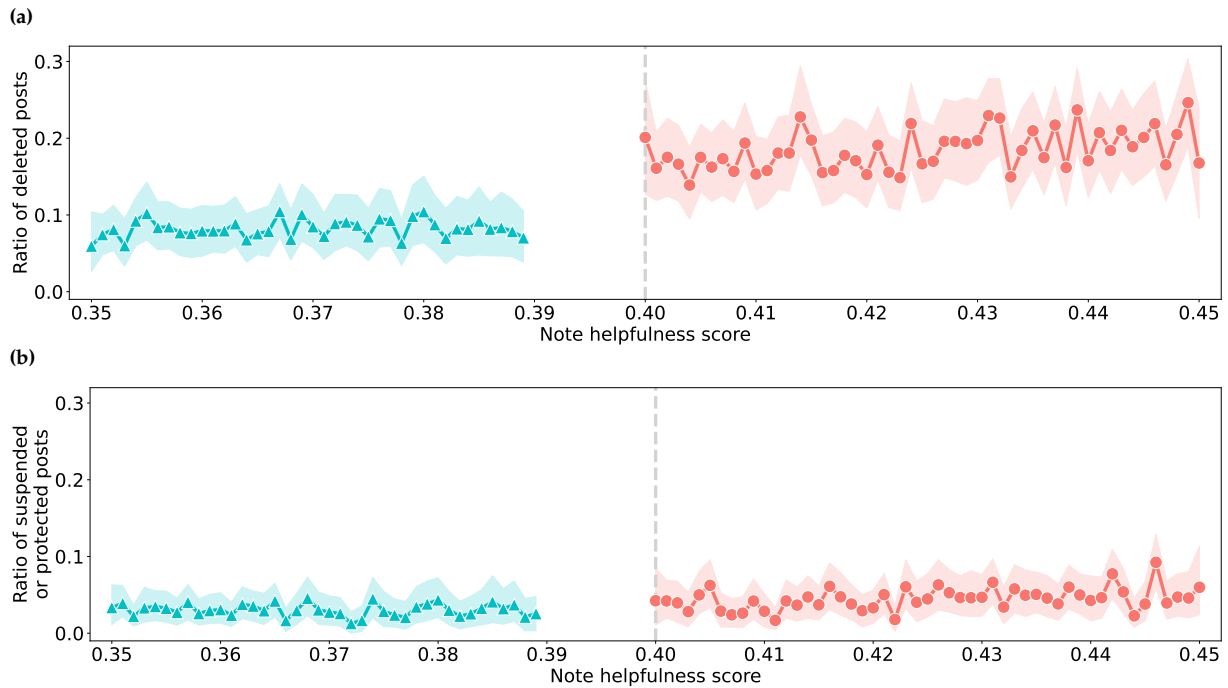


Figure 6.5: Effect of community notes on post deletion and suspension/protection. (a) Shown are the ratios of deleted posts for different values of note helpfulness. (b) Shown are the ratios of protected/suspended posts for different values of note helpfulness. Note helpfulness scores are re-centered based on the cutoff point of 0.40. Only community notes with note helpfulness scores of 0.40 or above are displayed on the corresponding misleading posts. The note helpfulness scores are rounded to three decimal places. Notes with helpfulness scores of 0.39 are omitted to prevent treatment contamination due to fluctuations between recalculated note helpfulness scores and note helpfulness scores used in production. The error bands represent 99% CIs. See Suppl. S6.10 for further details and full estimation results.

of the note helpfulness score, and the treatment effect is estimated by comparing the outcomes just above and just below the cutoff point.

The estimated changes in the ratio of deleted posts for different note helpfulness scores are visualized in Fig. 6.5a. We observed a clear sharp discontinuity at the cut-off point of 0.40. Specifically, the estimated treatment effect of community notes display in our RDD model was 0.943 (99% CI: [0.611, 1.342]; $p < 0.001$), which implies that the odds for deletion by their authors were 94.3% higher for posts with displayed community notes than for posts without displayed community notes (see Suppl. S6.10).

In addition to deletion, misleading posts with displayed notes could become inaccessible either because they were protected by their authors (*i.e.*, private posts) or due to account suspensions imposed by the $\mathbb{X}$ platform. Although the $\mathbb{X}$ API does not distinguish between these two scenarios, a manual review of 1500 posts revealed that almost all (94.4%) of these inaccessible posts resulted from account suspensions (*i.e.*, platform-enforced actions). We adopted the same RDD approach to examine the potential effect of community notes on post suspension/protection (see Suppl. S6.10 for details). Unlike the findings for post deletions, Fig. 6.5b shows that there was no sharp change in the ratio of suspended/protected posts around the cut-off point of note helpfulness scores. Furthermore, the coefficient estimate for the treatment effect of community notes on post suspension/protection was not statistically significant (see Suppl. S6.10), reinforcing this observation. In summary, our analysis provides confirmatory evidence that community notes did not trigger enforcement of other content moderation rules on $\mathbb{X}$, *i.e.*, platform-enforced actions (see Suppl. S6.11 for details).

6.3 Discussion

We performed a large-scale quasi-experimental study to assess the real-world efficacy of community-based fact-checking on the social media platform $\mathbb{X}$. Based on time series data for $N = 237,180$ (community fact-checked) $\mathbb{X}$ cascades that had been reposted more than 431 million times, we found that exposing users to community notes reduced the spread of misleading posts by, on average, 61.2%. This is largely consistent with small-scale A/B tests carried out by $\mathbb{X}$ internally during the pilot phase of community notes [67]. Furthermore, the odds for post deletion were 94.3% higher for posts with displayed community notes than for posts without displayed community notes. Crucially, $\mathbb{X}$'s recommendation algorithm (available as open source [295]) does not impose any penalties (*e.g.*, visibility reduction) on posts flagged with community notes [296]. This implies that the observed effects must be attributed to shifts in user behavior, rather than algorithmic intervention (Suppl. S6.11). Altogether, our findings complement earlier studies focusing on the question of whether crowds are able to accurately assess social media content [46, 47, 51–53, 279–281], by offering real-world empirical evidence that community notes reduces sharing of misleading posts on $\mathbb{X}$.

The treatment effect of community notes was pronounced and statistically significant across the board. Individuals across both sides of the political spectrum, even those who are frequent consumers of misinformation, were less likely to share misleading posts when informed by community notes. Notwithstanding, we still observed substantial variations concerning both user characteristics and content characteristics. Specifically, the efficacy of community-based fact-checking was discounted for posts originating from influential accounts (*e.g.*, verified users, high-follower accounts). A potential explanation is that such users may have the ability to attract more like-minded individuals, fostering echo chambers where their false beliefs are reinforced [17], thereby potentially mitigating the efficacy of community notes. Consistent with this notion, our results indicate that a primary mechanism through which community fact-checks reduce engagement is by discouraging resharing among users who had no prior interaction with the author of the misleading posts (*i.e.*, disconnected users). Put differently, this implies that community fact-checks may be less effective in tackling misinformation disseminated by influential accounts with dedicated and trusting audiences (*e.g.*, politicians, public figures). In addition to user heterogeneity, content characteristics also significantly influenced the efficacy of community notes. The presence of attached media items (*e.g.*, images, videos) significantly amplified the intervention effect. In contrast, community notes were least efficacious when addressing health-related or political misinformation.

While community notes exhibited a pronounced treatment effect, the timing of their display poses a significant challenge. Community-based fact-checking tends to be faster than professional fact-checking [286], yet the rapid dissemination of posts on $\mathbb{X}$ still outpaced the (relatively) slow display of community notes, limiting the feature's ability to curb misinformation early on. In our data, the half-life of reposts over 36 hours is 5.75 hours, with only 13.5% of all helpful notes displayed before this time point. Such delays reduced their effectiveness during the early, and often most viral, stages of diffusion. As a result, their system-wide effect was more modest, lowering the total number of reposts of misleading posts by 14.9%. This effect was smaller during the early phases of the Community Notes program, consistent with Chuai et al. [74] who analyzed aggregated engagement metrics from before

and shortly after the feature's rollout (see Suppl. S6.9.4), but has grown stronger over time. Notably, the observed effect sizes are also broadly in line with earlier evidence on the efficacy of fact-checking interventions. Internal data from Meta suggested that labels on false election posts reduced sharing by $\approx 8\%$ cumulatively, whereas the company's self-reported labeling and downranking of fact-checked false content was claimed to reduce subsequent sharing by $\approx 80\%$ [297]. By comparison, community notes achieved a 61.2% decrease in subsequent sharing, even without downranking measures. Altogether, the key implication is thus that community notes are effective in reducing the spread of posts once annotated, but their system-wide impact depends critically on speed: notes must appear early, ideally before misleading posts gain traction and spread widely.

As with any study, ours is not without limitations, which present opportunities for future research. First, our focus on community-based fact-checking was limited to one social media platform, specifically examining the effectiveness of community notes on X. While this platform represents the only large-scale implementation of community-based fact-checking on mainstream social media platforms to date, future research could expand to include other platforms if they choose to implement similar features (*e.g.*, YouTube and Meta have recently begun testing a fact-checking feature comparable to X's Community Notes). Second, the Community Notes feature itself may evolve over time, potentially influencing its effectiveness in combating misinformation. While we took intensive care to account for potential effects of algorithmic changes in our empirical analysis (see Suppl. S6.11), future research should examine how further developments or enhancements to this feature influence its efficacy. Third, incorporating information on network structure could yield additional insights. For example, recent research [298] confirms our findings and shows that note attachment qualitatively changes the structure of the post's diffusion cascade. Fourth, our study primarily focused on the American context and the global North, with a specific emphasis on posts in English. Future research could explore the efficacy of community-based fact-checking across different cultural contexts to gain a more comprehensive understanding of its impact on diverse communities. Finally, it could be valuable to investigate how the intervention effect varies across different demographic groups, such as age, gender, or education level, to uncover insights into who is most influenced by community-based fact-checking efforts.

Fact-checking becomes increasingly vital, particularly in light of emerging challenges posed by the scalability of AI-generated misinformation. Countermeasures against misinformation are more relevant than ever, and community-based interventions, such as community notes, hold promise in this fight, at least in concept. However, the effectiveness of such interventions hinges on their ability to respond swiftly to the dynamic nature of misinformation dissemination. Therefore, enhancing the speed and efficacy of community-based fact-checking approaches is crucial in fortifying our defenses against the spread of false information in an era marked by heightened digital influence.

6.4 Methods

6.4.1 Data Collection

We downloaded all community note contributions via a dedicated website on July 14, 2024 [299]. X routinely updates the community note contributions on a daily base and releases them to the public via this website. The community note contributions were documented in three distinct datasets: notes, ratings, and note status history. Each note in the notes dataset includes its unique note ID, referenced post ID, creation timestamp, and classification (*i.e.*, “misinformed or potentially misleading” or “not misleading”). The ratings dataset includes all the ratings submitted by the contributors along with their creation timestamps. The note status history dataset records the changes in the status of referenced notes and corresponding timestamps. The recorded note statuses contain the information on the helpfulness ratings, *i.e.*, whether the notes were displayed on the corresponding misleading posts.

Based on the referenced post IDs in the notes dataset, we collected the corresponding source posts via the X API v2 post lookup endpoint. Here, we only considered source posts written in English. Each post object returned from this endpoint contains detailed information related to the post itself (such as text, creation time, and media metadata) and its author profile (such as verified status, number of followers, and number of followees). Subsequently, we employed the X API v2 full-archive post counts endpoint to collect *time series data* of repost counts at minute level for the community fact-checked posts over a period of 36 hours since their creation. In total, our dataset contains time series data of repost counts for a total of $N = 237,180$ (community fact-checked) posts that were created between October 6, 2022 and June 11, 2024 and had been reshared more than 431 million times on X. A detailed overview of the dataset and descriptive statistics are available in Suppl. S6.1 and Suppl. S6.2, respectively.

6.4.2 Empirical Analysis

We employed a Difference-in-Differences design to estimate the treatment effect of community notes on the spread of misleading posts on X as measured by the repost count over time following the note display. The treatment group consists of all source posts with displayed notes, *i.e.*, posts that received a community note rated as helpful and, thus, displayed to all users on X. We considered the time of the display of the first helpful note for each post as the start of the treatment.

To ensure comparability, we constructed a control group from the source posts that share similar characteristics with the source posts in the treatment group but have no displayed notes. Given that the number of source posts without displayed notes (201,098) is nearly 6 times the number of source posts with displayed notes (36,082), we performed one-to-one propensity score matching for source posts with displayed notes and find their closest matches in the source post without displayed notes to construct a balanced control group (see Suppl. S6.4). The one-to-one matching was conducted based on the variables from user profiles (*e.g.*, followers and followees) and post features (*e.g.*, sentiments and topics). As a result, the source posts in the control group have no statistically significant differences from the source posts in the treatment group with respect to the user and post characteristics. Subsequently, we

assigned the virtual time of note display to the source posts in the control group according to the corresponding posts in the treatment group and recentered the repost timelines around the display of community notes.

Two-period DiD: We implemented a two-period DiD model to estimate differences in repost counts between the treatment and control group before and after the display of community notes. In our data, community notes were displayed 1.239 hours after post creation at the earliest, and only 5.2% of community notes were displayed within 4 hours since post creation. Therefore, we considered a 4-hour window before the note display as before-display period. The after-display period was set to between 1 and 12 hours from the note display. The dependent variable in the DiD model is the number of reposts (*i.e.*, *RepostCount*). Given that *RepostCount* is a non-negative count variable with overdispersion, we employed a multilevel mixed-effects negative binomial regression. Formally, we specified the negative binomial regression model with post-level random effects (*i.e.*, random intercepts) for the two-period ATT estimation as:

$$\begin{aligned} \log(E(\text{RepostCount}_{it}|x_{it})) = & \beta_0 + \beta_1 \text{Display}_i + \beta_2 \text{After}_t \\ & + \beta_3 \text{Display}_i \times \text{After}_t \\ & + \beta_4 \text{PostAge}_{it} + \mu_{post}, \end{aligned} \quad (6.1)$$

where x_{it} indicates all the independent variables and β_0 is the intercept. Display_i is a group-specific dummy variable indicating whether the source post i belongs to the treatment group ($= 1$) or control group ($= 0$). After_t is a time-specific dummy variable indicating the time is before ($= 0$) or after the note display ($= 1$). $\text{Display}_i \times \text{After}_t$ is the difference-in-differences interaction capturing the treatment effect. The variable μ_{post} represents post-specific random effects, which capture post-level heterogeneity that cannot be reflected by the general post features in the propensity score matching. For example, novel posts related to real-world events that happened suddenly may gain more engagement.

In our main analysis, we opted to use post-level random effects rather than fixed effects for the following reasons: (i) We conducted one-to-one propensity score matching to construct a control group that was balanced with the treatment group across all poster and post characteristics (see Suppl. S6.4). This effectively mitigates concerns that unobserved poster-specific or post-specific factors could be related to covariates. (ii) Our estimation was based on a large scale of 654,400 longitudinal observations across $N = 40,900$ posts, where the assumption of normal distribution of post-level random effects was appropriate. (iii) Fixed-effects models would omit posts with constant outcomes, *i.e.*, repost counts, over time. Given that the distribution of repost counts were overdispersed with many observations at zero, dropping these constant observations would compromise the representativeness of our dataset. Nonetheless, we also repeated our analysis using post-level fixed effects, where we observed consistent results (see Suppl. S6.9.1). Additionally, to account for the possibility that excess zeros in repost counts at later stages of diffusion, we repeated our analysis with zero-inflated negative binomial regression models. Also here, the results were robust and consistent with our main analysis (see Suppl. S6.9.2).

Multi-period DiD: We estimated a multi-period DiD to enable a temporal analysis of the treatment effect. Here, we again considered a 4-hour window before the note display as the before-display period and then examined the hourly multi-period ATTs from 1 to 12 hours after the note

display. The baseline of time effects is the first hour before note display. Formally, we specified the negative binomial regression model for the multi-period ATT estimation as:

$$\begin{aligned} \log(E(\text{RepostCount}_{it}|x_{it})) = & \beta_0 + \beta_1 \text{Display}_i + \beta'_2 \text{Before}_t + \beta'_3 \text{After}_t \\ & + \beta'_4 \text{Display}_i \times \text{Before}_t \\ & + \beta'_5 \text{Display}_i \times \text{After}_t \\ & + \beta_6 \text{PostAge}_{it} + \mu_{post}, \end{aligned} \quad (6.2)$$

where β_4 is the estimates for the parallel test, and β_5 is the estimates for the corresponding difference-in-difference interactions that can be transformed to multi-period ATTs. Further, PostAge_{it} denotes the age of source post i at time t since its creation. All other variables are the same as those in the two-period model.

Calculation of ATTs: The coefficients estimated for the difference-in-differences terms in the two-period and multi-period regression models are the natural logarithms of the ratios for the number of reposts with the treatment of note display compared to the number of reposts that are expected to receive without the note display during the after-display period. We exponentially transformed the coefficient estimates of the difference-in-differences terms in the models and measured the treatment effect of community notes (*i.e.*, ATT) as:

$$ATT = e^\beta - 1, \quad (6.3)$$

where β is the coefficient estimates for the difference-in-differences terms. ATT indicates the ratio of extra change of the reposts in the source posts with displayed notes relative to the reposts that the source posts are expected to receive without displayed notes. We used this indicator to analyze the efficacy of community notes.

Moral Outrage in Replies

Abstract: Displaying community fact-checks is a promising approach to reduce engagement with misinformation on social media. However, how users respond to misleading content *emotionally* after community fact-checks are displayed on posts is unclear. Here, we employ quasi-experimental methods to *causally* analyze changes in sentiments and (moral) emotions in replies to misleading posts following the display of community fact-checks. Our evaluation is based on a large-scale panel dataset comprising $N = 2,225,260$ replies across 1841 source posts from X’s Community Notes platform. We find that informing users about falsehoods through community fact-checks significantly increases negativity (by 7.3%), anger (by 13.2%), disgust (by 4.7%), and moral outrage (by 16.0%) in the corresponding replies. These results indicate that users perceive spreading misinformation as a violation of social norms and that those who spread misinformation should expect negative reactions once their content is debunked. We derive important implications for the design of community-based fact-checking systems.

Publication

Y. Chuai, A. Sergeeva, G. Lenzini, and N. Pröllochs, ‘Community fact-checks trigger moral outrage in replies to misleading posts on social media,’ in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025, pp. 1–23.

7.1 Introduction

Social media platforms like X (formerly Twitter), Facebook, and TikTok have become increasingly influential sources for news consumption and social engagement, especially among younger generations [300]. However, the fast-paced and interactive nature of social media has also compromised the quality of the information shared, as content can be posted rapidly without undergoing the thorough fact-checking essential to journalistic ethics [301]. As a result, misinformation spreads easily on social media platforms and has become a concerning issue in the digital age [5, 14, 302]. The negative effects of user engagement with online misinformation have been repeatedly observed across various domains [1–4, 21, 33]. For instance, political misinformation is especially troubling, as engagement with false information during elections can undermine democracy and destabilize societies [4, 14, 33, 253]. Given the significant risks and threats posed by online misinformation, social media providers are thus urged to establish content moderation policies and implement effective interventions to identify misinformation and inform users on their platforms [5, 141, 263, 303].

As a countermeasure, the social media platform X has recently adopted a crowdsourced (community-based) fact-checking approach, “Community Notes,” to fact-check misleading posts and communicate veracity to users on its platform [167]. “Community Notes” represents the first attempt by a major social media platform to apply a crowdsourced approach to fact-checking on a large scale. It allows users to add context

7.1	Introduction	119
7.2	Background and Related Work	122
7.2.1	Misinformation on Social Media	122
7.2.2	Content Moderation for Misinformation	123
7.2.3	Sentiments, Basic Emotions, and Moral Outrage on Social Media	123
7.3	Data and Methods	125
7.3.1	Datasets	125
7.3.2	Sentiments and Basic Emotions	126
7.3.3	Political vs. Non-Political Posts	127
7.3.4	Empirical Models	128
7.4	Results	131
7.4.1	Data Overview	131
7.4.2	Analysis of Sentiments and Basic Emotions (RQ1.1)	134
7.4.3	Analysis Across Political and Non-Political Misleading Posts (RQ1.2)	138
7.4.4	Analysis of Moral Outrage (RQ2.1 & RQ2.2)	141
7.4.5	Robustness Checks	143
7.4.6	Summary of Main Findings	144
7.5	Discussion	145
7.5.1	Relevance	145
7.5.2	Interpretations	145
7.5.3	Implications for Further Design of Community-Based Fact-Checking	147
7.5.4	Limitations and Future Research	148
7.6	Conclusion	149

or correct information to posts via *community notes*, providing additional perspectives or clarifications. Once community notes are rated helpful by other community contributors, they will be directly displayed on the corresponding misleading posts. While the accuracy of crowdsourced fact-checking approaches has been thoroughly validated by previous research [51–54, 96], a causal understanding of how users react to social media content that has been subject to community-based fact-checking is still in its infancy.

Previous research studying the effect of community fact-checks on social media users has primarily focused on analyzing how it affects *engagement* with misinformation (*e.g.*, shares, comments, replies, or likes) [56, 74, 76, 96, 304]. For instance, studies have shown that community notes can reduce users' belief in false content and their intentions to share misleading posts [76, 179, 304]. However, little is known about how social media users *emotionally* react to social media posts that have been subject to fact-checking. Emotions are highly influential for human decision-making and judgment [305]. It is thus conceivable that fact-checking social media posts triggers emotional backlash towards the authors of such content.¹ For example, users may feel a sense of urgency to correct the record or call out the spreaders of misinformation. In this scenario, the debunking of misinformation via community notes might provoke strong emotional reactions (*e.g.*, negative sentiment, anger, disgust) from social media users. Note that such emotional backlash from the user base may be seen as a double-edged sword. On the one hand, it can be a desirable reaction in the fight against misinformation as it may signal to the content creator and other users that spreading misinformation is unacceptable and subject to public scrutiny. Also, collective emotional backlash may help pressure authors to correct or delete misleading content and discourage others from sharing or engaging with it. On the other hand, however, social media platforms also need to maintain a healthy and constructive dialogue. In this regard, the emotional reactions from the user base pose risks of fueling political polarization, escalating conflicts, and entrenching divisions. Yet, empirical evidence on how fact-checking emotionally resonates with the audience on social media is missing.

1: Note that an alternative rationale posits that some users (*e.g.*, those with strong partisan alignments) may respond emotionally to a post from their "side" being labeled as misleading, rather than reacting to the misinformation itself. However, previous research suggests that replies are typically aimed at the content or the post authors [306]; and users are more likely to reply to authors from ideologically opposing groups [307]. In line with this notion, our later analysis and manual validation imply that community notes primarily elicit replies focused on the content of misleading posts or their authors (see details in Section 7.4.4).

Furthermore, fact-checking social media posts may trigger pronounced moral reactions from the user base. Content moderation of online speech is a moral minefield, especially when two key values come into conflict: upholding freedom of expression and preventing harm caused by the violation of social norms and spreading misinformation [308]. From a theoretical perspective, spreading online misinformation may be perceived as immoral behavior, motivating users' moral reasoning and judgment and triggering their moral emotions [309]. Previous work has identified the phenomenon of *moral outrage*, which occurs when individuals perceive the actions of another person or an entire situation as a serious violation of their moral principles or ethical standards. Moral outrage is commonly associated with a higher likelihood of inter-group hostility and can contribute to the further escalation of conflicts on the platforms [300, 309, 310]. Emotionally, this often results in a combination of anger and disgust, prompting individuals to express these emotions in an attempt to restore what they perceive as the proper moral norm [311]. This response typically involves blaming the violator and demanding corrective actions [300, 309]. Based on this rationale, we hypothesize that social media users might react with expressions of moral outrage to misleading posts and their authors once they are informed about the falsehoods via community notes.

Research questions: In this study, we *causally* analyze whether the display of community notes changes expressions of sentiments and (moral) emotions in social media users’ replies to misleading posts. Specifically, we address two main research questions. Our first research question is to examine the effects of community notes on sentiments and basic emotions in replies (RQ 1.1). Previous research further suggests that emotions may play a particularly pronounced role in the context of political (mis-)information [105, 183]. We thus additionally analyze how the emotional reactions differ across misleading posts that cover a political vs. a non-political topic (RQ1.2):

- **RQ1.1:** Which sentiments (positivity, negativity) and basic emotions (*e.g.*, anger, disgust) in replies to misleading posts are triggered by the display of community notes?
- **RQ1.2:** Do the effects of community notes on sentiments and basic emotions in replies differ across political and non-political misleading posts?

Furthermore, when anger and disgust are mixed, they generate “moral outrage,” a more intense emotional reaction to perceived moral transgressions than either anger or disgust alone [310–312]. Our second research question is to examine whether community notes trigger moral outrage in replies (RQ2.1), and analyze the sensitivity across political and non-political misleading posts (RQ2.2):

- **RQ2.1:** Does the display of community notes trigger moral outrage in replies to misleading posts?
- **RQ2.2:** Do the effects of community notes on moral outrage in replies differ across political and non-political misleading posts?

Data & methods: To address our research questions, we collect a dataset of $N = 2,225,260$ replies to 1841 misleading posts that have been fact-checked via community notes over a 4-month period since the roll-out of “Community Notes” feature on X/Twitter. The dataset includes all the replies that are directly toward misleading posts before and after the display of community notes. Subsequently, we employ methods from Natural Language Processing (NLP) in combination with Regression Discontinuity Design (RDD) to estimate the causal effects of community notes on the sentiments and (moral) emotions in the replies to the fact-checked posts.

Contributions: To the best of our knowledge, this study is the first to causally analyze the emotional reactions of social media users in response to (community) fact-checks. We find that informing users about falsehoods through the display of community fact-checks significantly increases the expression of negativity in the corresponding replies by 7.3%. Furthermore, we provide a fine-granular analysis of emotions. Here, the largest effect sizes are observed for (negative) emotions from the moral emotions family. Specifically, displaying community notes increases anger in the corresponding replies by 13.2%, disgust by 4.7%, and moral outrage by 16.0%. The observed effects are significantly more pronounced for political vs. non-political posts. Overall, these results imply that social media users perceive misinformation as moral transgressions and that authors of misinformation should expect negative reactions once their content is debunked. Based on our findings, we derive important implications for behavioral theory, media literacy, and the design of community-based fact-checking systems on social media.

7.2 Background and Related Work

7.2.1 Misinformation on Social Media

Social media platforms, as a source for news consumption and social engagement, have become increasingly popular. As of July 2024, there is a large population base of ~5.17 billion users on social media, and the number continues to grow [300, 313]. This indicates a trend toward partially substituting traditional media with social media channels that provide fast access to information [314] and offer a more interactive way to engage with content [315].

Due to the lack of editorial oversight, social media platforms have also become fertile soil for spreading misinformation. Recently, there have been some critics arguing that the prevalence of misinformation on social media is overestimated, and the exposure is concentrated among narrow fringe groups [30, 31]. However, given the large population base on social media platforms, its negative effects can be profound. For example, the spread of online health misinformation has posed severe risks to public health in both the Global North and the Global South [2, 3, 28, 316]. Additionally, the rise of AI makes generating misinformation faster, which poses new challenges and further undermines the credibility of social media [155, 317–319].

The spread of political misinformation on social media: The spread of political misinformation on social media has become a tool to pursue political agenda, and its negative effects are continually observed in threatening the stability of societies and eroding trust in democracy [1, 21, 33]. For example, conspiracy theories claiming that the 2020 U.S. presidential election was “stolen” were widely shared on social media. It is found that users who were exposed to at least one misinformation website are 17.3% more likely to believe that the election was fraudulent [320]. As surveyed, almost 40% of the U.S. public and, especially, almost 70% of Republicans negated the legitimacy of the 2020 presidential election outcome [321]. The reduced faith in electoral institutions have had real-world consequences in affecting voters’ behaviors in subsequent elections and hate campaigns [4, 33]. Similar patterns have been observed in countries of the Global South, such as in Brazil [19, 20]. Moreover, political misinformation is often interconnected with the politicization of other social issues, such as health crises and climate change, further contributing to polarized attitudes toward these topics [72, 73, 121].

Altogether, with the increasing use of social media for news consumption and the continuously observed negative effects of misinformation across various domains, the spread of misinformation on social media has become a significant societal issue in the digital age [14, 300]. Particularly, political misinformation has been found to spread significantly faster than non-political misinformation on social media [14]. Previous studies also show that social media users share information to single their political affiliation and moral stance, even knowing that it is inaccurate [102, 322]. To protect users from being affected by misinformation and prevent harm, it is both warranted and required to implement effective countermeasures to fight against its spread [33, 253].

7.2.2 Content Moderation for Misinformation

Social media providers are increasingly urged to establish content moderation policies to regulate online behavior and combat the spread of misinformation on their platforms [5, 263]. The content moderation process includes two main aspects: fact-checking misinformation and handling identified misinformation.

To identify misinformation, social media platforms are widely collaborating with professional third-party fact-checkers (*e.g.*, the International Fact-Checking Network) [176]. While fact-checking experts can accurately access veracity [152, 156], the implementation of this approach faces two main challenges. First, professional fact-checking is a lengthy and resource-consuming process, leading to a situation where only a small proportion of posts are actually checked [150, 156, 323, 324]. Second, professional fact-checkers face suspicions of (political) bias, *e.g.*, by selectively fact-checking content [73, 325]. Therefore, the trust of the public in professional fact-checking is discounted [56, 76, 278]. To overcome these challenges, researchers have proposed to fact-check and identify misleading posts on social media platforms at scale based on the non-expert fact-checkers from the crowd [46, 53, 279]. While the individual assessments can have bias and noise [50], it has been shown that the accuracy of aggregated judgments, even from relatively small crowds, is reliable and comparable to the accuracy of expert fact-checkers [51–54]. Additionally, the crowdsourced assessments are perceived as more trustworthy by the users than fact-checks from experts [46, 55, 56].

Subsequently, social media platforms generally have three options for handling identified misleading posts: removing them, reducing their visibility, or informing users about the falsehoods [67, 176, 326]. The first two options – removal and visibility reduction – are forms of “hard moderation,” where the platforms limit or entirely cut off the spread of misinformation. In contrast, informing other users that a post is misleading represents “soft moderation,” where the platforms refrain from imposing direct controls and instead allow users to make their own decisions on how they process the informed misleading post. The hard moderation approaches face a dilemma, where the two key values of protecting freedom of expression and preventing harm from the spread of misinformation come into conflict [308, 327]. Given this, soft moderation approaches can be regarded as a trade-off between freedom of expression and misinformation intervention. Previous lab/survey experiments consistently show that communicating the falsehood to users (especially with explanations) is effective in reducing users’ beliefs and sharing intentions [46, 55, 190, 270, 274, 328–330].

7.2.3 Sentiments, Basic Emotions, and Moral Outrage on Social Media

As an essential element of human communication, sentiments and emotions play a pivotal role in shaping interactions on social media platforms [305, 331–333]. They can be understood as two affective dimensions: valence (sentiment) and arousal (emotion). Sentiment, *i.e.*, the valence of content (positive vs. negative), is an attitude, thought, or judgment promoted by a feeling [334]. Research has shown that negativity is a particularly significant characteristic of online engagement [98, 335, 336]. Negative sentiment tends to evoke stronger reactions, promoting

users to comment and share online posts, especially among polarized communities [335, 336].

The relationship between sentiments and social interaction is more nuanced than simply the valence (positive vs. negative) of the content alone. From the arousal dimension, specific emotions (*i.e.*, affect ladenness) are embedded within the content, adding intensity to the positive or negative feelings expressed [334, 337]. With various models classifying basic emotions, Ekman's six basic emotion model is widely used and commonly shared across other frameworks [334]. Specifically, Ekman's model identifies six universal emotions: anger, disgust, fear, joy, sadness, and surprise [338–340]. These emotions are considered fundamental across different cultures and can combine to form more complex emotional experiences [339]. Emotions also vary on the level of physiological arousal or activation they evoke [332]. For example, anger and sadness are often expressed as negative emotions. However, anger is a state of heightened arousal and makes online content more viral, while sadness is low arousal and makes online content less viral [82, 332]. Particularly, users' reliance on emotions can promote their belief in misinformation [341]. Several studies have found that emotions can explain the spread of true vs. false information on social media, with anger being the most significant emotion in driving the spread of misinformation [10, 34].

Additionally, *moral emotions* can arise in response to perceived moral violations or standards and motivate moral behavior [342]. One large family of moral emotions is other-condemning emotions that often lead to behavioral outcomes, that is, actions to correct the other's behavior and increase justice-relevant judgment [342, 343]. Specifically, anger and disgust are the main other-condemning emotions used in coordinating and constraining the behavior of the group member who made the morally unacceptable actions [342]. During the COVID-19 pandemic, other-condemning emotions are significantly associated with the higher virality of misinformation on social media [81]. Within political contexts, social media posts containing moral-emotional language have been shown to be 20% more likely to be shared, particularly within polarized partisan networks (liberal or conservative), with stronger effects within these groups than between them [183]. This highlights the role of moral emotions in amplifying online engagement, particularly through expressions of out-group animosity. Such expressions, especially within partisan networks, have been shown to drive higher engagement than in-group favoritism, further entrenching political polarization [105].

Previous studies highlight *moral outrage* – a mixture of anger and disgust triggered by a perceived moral norm violation – as a powerful emotion that motivates people to shame and punish wrongdoers, especially in the digital age [310, 311, 344, 345]. Social media platforms, such as X/Twitter, inherently amplify expressions of moral outrage over time, as users who learn such language get rewarded with an increased number of likes and reposts [312]. Moral outrage is also contagious. Knowing that a person or entity is being publicly condemned can provoke moral outrage in others [346]. This contagious has the potential to ignite online firestorms – large waves of insults, criticism, or swearing directed at persons and organizations on social media [347, 348]. In addition, users frequently overestimate the level of outrage and hostility in social media posts, perceiving emotional expressions as more extreme than the authors actually report. This overperception can lead to escalated responses and increase the spread of moral outrage [344]. The consequences of heightened moral outrage on social media are double-edged. On the one hand, it can fuel inter-group conflict and polarization, exacerbating online

tensions and divisions. On the other hand, it can also amplify constructive facets of morality, such as social support and prosociality [300, 310]. This dual role underscores the importance of understanding how moral outrage operates in specific circumstances to mitigate its negative impacts while promoting its potential positive effects.

Overall, the analysis of sentiments and (moral) emotions is a crucial and growing area of research for understanding online engagement and its consequences [162]. A large research body has documented that sentiments, (moral) emotions, and moral outrage play a pronounced role in shaping engagement with (mis-)information on social media. However, an understanding of how users emotionally react to social media posts after they are informed of the falsehoods is missing, which presents the aim of our study.

7.3 Data and Methods

7.3.1 Datasets

In this study, we analyze the emotional reactions of social media users in response to (community) fact-checks. For this purpose, we collect a large dataset of posts that have been fact-checked on X's (formerly Twitter) "Community Notes" platform (formerly Birdwatch). "Community Notes" was officially rolled out to users on December 11, 2022, and is a crowd-sourced fact-checking program that allows users to identify posts that they believe are misleading and write short textual notes that provide context to the posts. After a community note is created, other users can rate its helpfulness and, if it reaches a certain level of helpfulness, it is displayed prominently beneath the original post. An example of a misleading post with a displayed community note and one direct reply is shown in Fig. 7.1a. Importantly, before note display, users who reply to the source posts can only see the original post content (as shown in Fig. 7.1b). After note display, users who reply to the source posts can see both original post content and displayed community notes flagging the post content as being misleading.

As detailed in the following, we collect data from three different sources: (i) note dataset, (ii) post dataset, and (iii) reply dataset. Subsequently, we combine these three data sources into a longitudinal dataset that allows us to analyze the changes in reply emotions before and after the display of community notes.

Note dataset. X publishes all community notes and their associated status histories on a dedicated website and updates them daily.² We downloaded the community note datasets on April 8, 2023, which include all note contributions up until April 6, 2023. Our note dataset contains the details for each note, including its unique ID, creation time, classification (misleading or not misleading), referenced source post ID, and the note status history (*i.e.*, when the notes were rated helpful and displayed on the source posts). In our data, there are 68,235 community notes for 45,664 source posts. However, only 4686 source posts (10.3%) received displayed community notes until the date of data collection.

Post dataset. We use X's Academic Research API to search for the 4686 source posts with displayed community notes and successfully collect 3872 (82.6%) posts. Notably, given that the "Community Notes" program was rolled out on December 11, 2022, and mainly developed in the U.S.,

2: All community notes are publicly available at <https://x.com/i/communitynotes/download-data>.

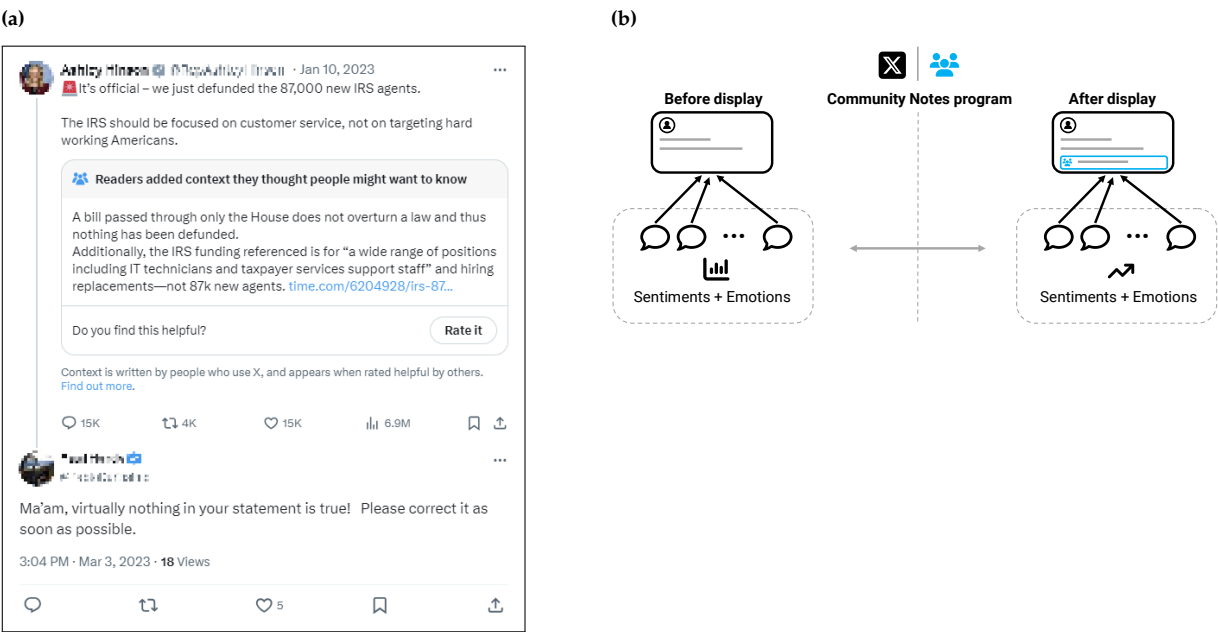


Figure 7.1: Research overview. (a) An example of a misleading post with a displayed community note and one direct reply after note display. (b) Illustration of our research setup. Before note display, users who reply to the source posts can only see the original post content. After note display, users who reply to the source posts can see both original post content and displayed community notes flagging the post content as misleading. Using the sentiments and emotions in replies before note display as a baseline, we can examine the changes in sentiments and emotions in replies after the display of community notes.

3: See the documentation for X’s API at <https://developer.x.com/en/docs/twitter-api/conversation-id>.

4: The collection of our dataset was completed over two months, from April to June 2023. Subsequently, X deprecated the Academic Research API, restricting our ability to access additional data for analysis. Notwithstanding, previous research implies that our dataset covering the initial four-month period provides a representative data basis for the Community Notes program [76, 304].

5: The sentiment model is available at Hugging Face.

6: The emotion model is available at Hugging Face.

we only consider the source posts that were created after the roll-out and are written in English. There are 1953 source posts remaining after this filtering step.

Reply dataset. We retrieve the entire conversation thread for each source post via X’s Academic Research API.³ Each conversation thread contains all the replies to the original posts, and they share one unique conversation ID. In this study, we analyze the changes of sentiments and emotions in replies to the original misleading posts after the display of community notes. Therefore, we only consider direct replies to the source posts and omit indirect replies (*i.e.*, replies to replies). We successfully pull $N = 2,225,260$ direct replies to 1841 misleading source posts (94.3%) that were fact-checked via community notes during a period of 4 months⁴ since its roll-out (Table 7.1). The remaining 5.7% (112) misleading source posts have no (English) replies and are thus omitted.

7.3.2 Sentiments and Basic Emotions

We extract sentiment and emotion features in the source posts and their replies using two state-of-the-art machine learning models. Specifically, we compute the probabilities of positive and negative sentiments for each source post and reply using the Twitter-roBERTa-base model (2022 updated). It was trained on 124 million posts created from January 2018 to December 2021, and fine-tuned for sentiment analysis.⁵ This model achieves a high prediction performance compared to other sentiment models [349, 350]. Subsequently, we compute the probabilities of Ekman’s six basic emotions (*i.e.*, anger, disgust, fear, joy, sadness, and surprise) [338, 339] for each source post and reply using a distilled version of the RoBERTa-base model.⁶ The model was pre-trained on OpenWebTextCorpus, a reproduction of OpenAI’s WebText dataset with

Table 7.1: Dataset overview. Reported are mean values or count numbers for the variables (standard deviations in parentheses).

	(1) All	(2) Political	(3) Non-political
#Replies	2,225,260	1,068,732	1,156,528
#Source posts	1841	477	1364
Date range of replies	2022-12-11 – 2023-05-28	2022-12-11 – 2023-05-27	2022-12-11 – 2023-05-28
Date range of source posts	2022-12-11 – 2023-04-05	2022-12-11 – 2023-04-05	2022-12-11 – 2023-04-05
<u>Sentiments in replies</u>			
Positive	0.112 (0.222)	0.098 (0.209)	0.124 (0.233)
Negative	0.565 (0.336)	0.608 (0.324)	0.525 (0.342)
<u>Sentiments in source posts</u>			
Source positive	0.114 (0.222)	0.084 (0.184)	0.125 (0.233)
Source negative	0.443 (0.334)	0.513 (0.315)	0.419 (0.337)
<u>Emotions in replies</u>			
Anger	0.181 (0.260)	0.205 (0.272)	0.160 (0.246)
Disgust	0.097 (0.183)	0.098 (0.181)	0.096 (0.185)
Fear	0.050 (0.146)	0.047 (0.138)	0.054 (0.153)
Joy	0.066 (0.170)	0.061 (0.163)	0.069 (0.176)
Sadness	0.107 (0.192)	0.109 (0.193)	0.104 (0.192)
Surprise	0.172 (0.247)	0.164 (0.241)	0.180 (0.252)
<u>Emotions in source posts</u>			
Source anger	0.159 (0.243)	0.195 (0.260)	0.147 (0.236)
Source disgust	0.041 (0.117)	0.041 (0.112)	0.040 (0.119)
Source fear	0.173 (0.276)	0.151 (0.242)	0.181 (0.287)
Source joy	0.077 (0.176)	0.068 (0.158)	0.080 (0.182)
Source sadness	0.126 (0.211)	0.140 (0.216)	0.122 (0.209)
Source surprise	0.175 (0.245)	0.174 (0.249)	0.175 (0.244)

more than 8 million documents, and fine-tuned for emotion analysis [351]. The performance of this emotion analysis model is comparable to human annotations and has been used in several previous studies [352, 353]. Summary statistics for the sentiments and emotions in the source posts and replies are reported in Table 7.1.

7.3.3 Political vs. Non-Political Posts

Misleading posts in the political domain are particularly concerning due to their potential to undermine democratic processes and destabilize societies [33]. As observed by previous studies [14, 102, 322], political misinformation spreads faster than non-political misinformation and is particularly resistant to corrections. Therefore, we additionally conduct a sensitivity analysis examining changes in sentiments and emotions in replies to misleading posts that cover a political vs. a non-political topic. To do so, we extract topic domains from the context annotations attached to each source post and classify them as political or non-political. X performs entity recognition and semantic analysis (e.g., keywords, hashtags, and handles) for all posts and assigns corresponding domain labels based on a topic domain list.⁷ Within the domain list, we consider “Politician,” “Political Body,” and “Political Race” as political topic domains. Posts containing one of the political topic domains are classified as political; otherwise, they are classified as non-political. As a result, there are 477 political misleading posts in our data, accounting for 25.9% of all misleading posts. However, nearly half of the replies are directed at

7: The topic domains are curated by domain experts and have undergone refinement over several years. X’s internal team audits this approach quarterly through the manual review of 10,000 posts across all the domains. The precision scores are consistently ~80%. The topic domain list is available at <https://developer.x.com/en/docs/twitter-api/annotations/overview>.

political misleading posts (see Table 7.1).

7.3.4 Empirical Models

In this study, we longitudinally examine the direct replies to the source posts before and after the display of community notes. To address our research questions, we use Regression Discontinuity Design (RDD) to causally estimate the effects of the display of community notes in terms of sentiments and basic emotions. RDD is a quasi-experimental method to estimate the average treatment effect of an intervention by exploiting a discontinuity in the relationship between a running variable and an outcome variable at a specific threshold or cutoff point [74, 184, 354, 355]. In the following, we introduce the dependent variables (Section 7.3.4), display indicator and running variable (Section 7.3.4), and additional independent variables (Section 7.3.4) in our analysis. Subsequently, we detail the specification of our empirical RDD models (Section 7.3.4).

Dependent variables.

We define the following dependent variables that measure expressions of sentiments, basic emotions, and moral outrage in replies to social media posts:

- ▶ Dependent variables for positive and negative sentiments
 - **Positive**: A continuous variable from 0 to 1 that indicates the probability of positive sentiment in a reply.
 - **Negative**: A continuous variable from 0 to 1 that indicates the probability of negative sentiment in a reply.
- ▶ Dependent variables for basic emotions
 - **Anger**: A continuous variable from 0 to 1 that indicates the probability of anger in a reply.
 - **Disgust**: A continuous variable from 0 to 1 that indicates the probability of disgust in a reply.
 - **Fear**: A continuous variable from 0 to 1 that indicates the probability of fear in a reply.
 - **Joy**: A continuous variable from 0 to 1 that indicates the probability of joy in a reply.
 - **Sadness**: A continuous variable from 0 to 1 that indicates the probability of sadness in a reply.
 - **Surprise**: A continuous variable from 0 to 1 that indicates the probability of surprise in a reply.
- ▶ Dependent variable for moral outrage
 - **MoralOutrage**: A continuous variable from 0 to 1 that indicates the probability of moral outrage in a reply. Moral outrage represents a unique combination of anger and disgust in which the two emotions interact so that each exacerbates the other's influence on judgments [311]. This interaction extends beyond the isolated effects of anger and disgust, thereby offering important theoretical insights into moral outrage [310–312]. Accordingly, we use the product of anger and disgust (*i.e.*, $Anger \times Disgust$) as a measure of moral outrage to explore their interactive effect.⁸

8: As a robustness check, we repeated our analysis with a dedicated classifier for digital outrage developed by previous research [312]. Here, we observe quantitatively identical results (see details in Suppl. S7.6).

Display indicator and running variable.

Based on RDD, we aim to examine whether there are significant discontinued changes in sentiments and emotions after the display of community notes and estimate the effects. The key rationale behind this is that the sentiments and emotions in replies should change smoothly (or keep stable) over time if without the display of community notes. Given this, we define the display indicator of community notes (binary treatment indicator) as *Displayed*. It is equal to 1 if a reply is created after note display; otherwise, it is equal to 0. Additionally, we examine the changes in sentiments and emotions over time before and after the display of community notes. Therefore, we define the running variable as *HoursFromDisplay*, which is a continuous variable indicating the hour(s) from note display to reply creation.

Additional independent variables.

Given that connected social media users may share similar emotions [356–358], the sentiments and emotions in the source posts are likely to affect the sentiments and emotions in the replies from users who are exposed to these source posts (see correlation analysis in Suppl. S7.1.1). Thus, we incorporate source sentiments and emotions into the regression models as independent variables:

- ▶ Independent variables for source sentiments in misleading posts
 - *SourcePositive*: A continuous variable from 0 to 1 that indicates the probability of positive sentiment in a source post.
 - *SourceNegative*: A continuous variable from 0 to 1 that indicates the probability of negative sentiment in a source post.
- ▶ Independent variables for source emotions in misleading posts
 - *SourceAnger*: A continuous variable from 0 to 1 that indicates the probability of anger in a source post.
 - *SourceDisgust*: A continuous variable from 0 to 1 that indicates the probability of disgust in a source post.
 - *SourceFear*: A continuous variable from 0 to 1 that indicates the probability of fear in a source post.
 - *SourceJoy*: A continuous variable from 0 to 1 that indicates the probability of joy in a source post.
 - *SourceSadness*: A continuous variable from 0 to 1 that indicates the probability of sadness in a source post.
 - *SourceSurprise*: A continuous variable from 0 to 1 that indicates the probability of surprise in a source post.

In addition, the times when community notes are displayed on the source post are not fixed and depend on note helpfulness scores that are recalculated every hour. Therefore, changes in sentiments and emotions after note display might be confounded by the post age from post creation to reply creation. Given this, we use *PostAge* to control for the possible confounding effects of post age. Descriptive statistics of sentiments and emotions in replies and source posts are summarized in Table 7.1.

Regression models.

We specify two RDD models to estimate the effects of the display of community notes on sentiments and emotions in replies, respectively.

For positive and negative sentiments, the RDD model is specified as:

$$\begin{aligned} \text{Sentiment}_i = & \beta_0 + \beta_1 \text{Displayed}_i + \beta_2 \text{HoursFromDisplay}_i \\ & + \beta_3 \text{SourcePositive}_i + \beta_4 \text{SourceNegative}_i \\ & + \beta_5 \text{PostAge}_i + \epsilon_i, \end{aligned} \quad (7.1)$$

where *Sentiment* denotes the dependent variable of *Positive* or *Negative*, β_0 is the intercept, and ϵ is the residual. β_1 to β_5 are coefficient estimates for the independent variables.

For the six basic emotions, the RDD model is specified as:

$$\begin{aligned} \text{Emotion}_i = & \beta_0 + \beta_1 \text{Displayed}_i + \beta_2 \text{HoursFromDisplay}_i \\ & + \beta_3 \text{SourceAnger}_i + \beta_4 \text{SourceDisgust}_i \\ & + \beta_5 \text{SourceFear}_i + \beta_6 \text{SourceJoy}_i \\ & + \beta_7 \text{SourceSadness}_i + \beta_8 \text{SourceSurprise}_i \\ & + \beta_9 \text{PostAge}_i + \epsilon_i, \end{aligned} \quad (7.2)$$

where *Emotion* denotes the dependent variable of one of the six basic emotions: *Anger*, *Disgust*, *Fear*, *Joy*, *Sadness*, or *Surprise*. Additionally, we further take *MoralOutrage* (i.e., $\text{Anger} \times \text{Disgust}$) as the dependent variable to explore the changes in moral outrage after the display of community notes. The independent variables for source sentiments are replaced with the variables for source emotions.

To balance the scales of variables, we further z-standardize all continuous variables when fitting the RDD models. The standardization is based on:

$$z_i = \frac{x_i - \mu}{\sigma}, \quad (7.3)$$

where x_i denotes original value for a continuous variable X , and z_i is the standardized value. μ and σ are mean and standard deviation of the variable X , respectively. Therefore, the coefficient estimates for *Displayed*, i.e., β_1 , in the models cannot be directly interpreted as the effects of note display. Given this, we define that Y_{before} is the outcome of a dependent variable before the display of community notes and Y_{after} is the outcome of a dependent variable after the display of community notes. Assuming that all the other independent variables are equal for the outcomes of Y_{before} and Y_{after} , we have:

$$\begin{aligned} Y_{\text{after}} - Y_{\text{before}} &= \beta_1 \\ \Rightarrow \frac{y_{\text{after}} - \mu}{\sigma} - \frac{y_{\text{before}} - \mu}{\sigma} &= \beta_1 \\ \Rightarrow y_{\text{after}} - y_{\text{before}} &= \sigma \beta_1. \end{aligned} \quad (7.4)$$

The value of $\sigma \beta_1$ can represent the estimated (predicted) effects of the display of community notes on the sentiments and emotions, with all else being equal.

Considerations for the adoption of RDD: Previous studies evaluating the effect of community notes on *engagement* (e.g., the number of reposts) have typically employed Difference-in-Differences (DiD) methods by constructing a counterfactual control group that evolves in parallel with the treatment group prior to the display of community notes [74, 76, 304]. However, constructing such control groups often omits a significant portion of posts. While our dataset includes a large volume of replies, the limited number of source posts makes it infeasible to conduct such matching at the post level. Additionally, previous studies focusing on

engagement typically analyze repost counts at an hourly level to examine engagement reductions attributed to the display of community notes. Therefore, they do not require the collection of every individual repost with a precise timestamp, and can count the number of reposts as zero for intervals without repost activity. In contrast, this paper focuses on analyzing sentiments and emotions in individual replies to misleading posts before and after the display of community notes. This analysis necessitates collecting every reply along with its associated timestamp and content. As sentiments and emotions cannot be assumed to be zero in intervals without replies, gaps in time units are likely to arise. These considerations lead us to adopt RDD with a continuous running variable in time [354, 355].

Further, research has discussed the use of time as a running variable in RDD, *i.e.*, Regression Discontinuity in Time (RDiT), and suggested considering the time-series nature of the underlying data-generating process [359, 360]. On social media platforms, the autoregressive time-series structure originates from the creation of original posts, extends to direct replies, and further encompasses indirect replies (*i.e.*, replies to replies), with each direct reply initiating a separate reply thread. To minimize complications from autoregressive processes within reply threads, our study focuses exclusively on direct replies to the original posts and omits indirect replies. Nevertheless, direct replies to the same source post across different observational units are still likely to exhibit correlations in terms of sentiments and emotions. However, according to X's algorithmic rules [361], replies in the conversations associated with the same source post are typically not displayed in chronological order. Thus, the correlations among direct replies are likely to primarily arise from shared factors, *e.g.*, source sentiments and emotions, in the original posts (as illustrated in Fig. 7.1b).⁹ Following prior research [14, 359], we cluster standard errors in all regression models by unique source posts to address the potential violation of independence assumption. By clustering the standard errors, we account for the potential within-cluster correlations and provide more precise and reliable standard error estimates. In addition to the cluster-robust method, the incorporation of source sentiments and emotions in the regression models explicitly controls for the correlations among replies linked to the same source posts.

9: In our later robustness checks, we validate that the correlations of sentiments and emotions among direct replies are stable and consistent across different temporal orders. The results confirm that the correlations arise from the influence of original posts rather than from temporal dependencies between direct replies. Furthermore, we repeated our analysis with lagged-dependent covariates as an additional check. Again, the results remain robust and support our findings (see details in Suppl. S7.7.2).

7.4 Results

7.4.1 Data Overview

Our dataset covers four months since the roll-out of the “Community Notes” program on X in December of 2022. Fig. 7.2a shows that the daily number of misleading posts on which community notes are displayed has an increasing trend, with an average of 15.7. Fig. 7.2b shows that misleading source posts tend to have higher negative sentiment (mean of 0.443), compared to positive sentiment (mean of 0.114; $t = -28.810$, $p < 0.001$). In terms of emotions, anger (mean of 0.159), fear (mean of 0.173), and surprise (mean of 0.175) are most common in the fact-checked misleading source posts (Fig. 7.2c).

To examine the effects of community notes on the sentiments and emotions in replies, we collect a large-scale longitudinal dataset of replies to misleading posts. Fig. 7.2d shows that each misleading source post

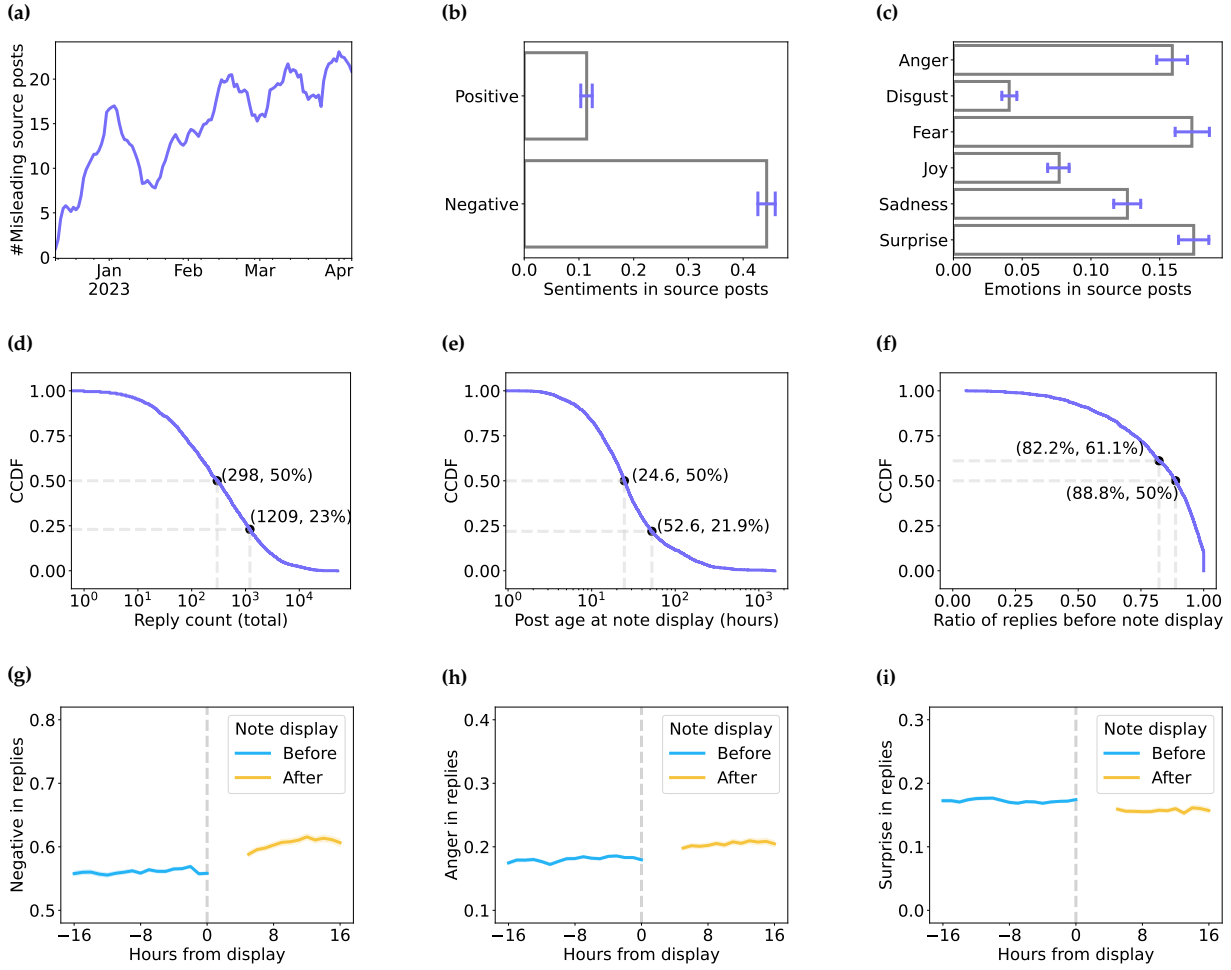


Figure 7.2: Summary statistics for misleading source posts, direct replies, and display of community notes. (a) The 2-week rolling average daily number of misleading source posts that are attached with displayed community notes. (b) The means of positive and negative sentiments in the misleading source posts. (c) The means of anger, disgust, fear, joy, sadness, and surprise in the misleading source posts. (d) The Complementary Cumulative Distribution Function (CCDF) for the number of replies that are directed at each misleading source post. (e) The CCDF for hour(s) from post creation to note display. (f) The CCDF for the ratio of replies before note display to the total replies. (g) The hourly averages of negative sentiment in replies across hours from note display. (h) The hourly averages of anger in replies across hours from note display. (i) The hourly averages of surprise in replies across hours from note display. The hourly averages of positive sentiment and other emotions are shown in Fig. S7.2 in Suppl. S7.1.2. The error bars (bands) represent 95% Confidence Intervals (CIs). Notably, previous studies consistently suggest a potential cold-start period of 4 hours for community notes to reach their full effect [76, 304]. Therefore, we visually omit reply points within the initial four hours after the display of community notes for better readability.

receives an average of 1209 (median of 298) replies in total (*i.e.*, before and after the display of community notes). On average, community notes are displayed 52.5 hours (median of 24.6 hours) after the creation of the source misleading posts (Fig. 7.2e). Notably, most of the replies (mean of 82.2%, median of 88.8%) are created before the display of community notes (Fig. 7.2f). Therefore, some misleading posts may not have sufficient reply samples after note display to perform comparisons with the replies before note display. Given this, we restrict our main analysis to misleading posts that have at least ($\geq$) 5 replies both before and after the display of community notes. This criterion¹⁰ results in 1339 misleading posts being included in the comparison of sentiments and emotions in replies before and after note display.

The key assumption of RDD in this paper is that a discontinuity exists around the cut-off point, *i.e.*, the time when community notes are dis-

10: This criterion is primarily applied to improve statistical inference. Our results remain robust and consistent when this restriction is removed (see details in Suppl. S7.7.4).

played [354, 355, 362]. In the absence of this intervention, the sentiments and emotions in replies are expected to remain continuous or stable. To validate this assumption, we first examine changes in sentiments and emotions over time before the display of community notes. Our analysis shows that, in the absence of community notes, the sentiments and emotions in replies remain relatively stable and exhibit no statistically significant relationship with the age of misleading posts (see details in Suppl. S7.1.2). Subsequently, we use an interrupted time series design to analyze potential discontinuities in sentiments and emotions in replies following the display of community notes. The results reveal statistically significant discontinuities for negative sentiment, anger, and surprise in replies after the display of community notes (each $p < 0.05$). Additionally, aside from the identified discontinuity at the cut-off point, the sentiments and emotions in replies after note display show trends similar to those in replies before note display and remain stable over time (see details in Suppl. S7.1.2). These findings confirm that RDD is appropriate for examining the effects of community notes on sentiments and emotions in replies.

Another important consideration for RDD is to include as many samples as possible within a relatively short bandwidth, aiming to mitigate potential confounding factors associated with larger windows around the display of community notes [354, 355, 359]. In our dataset, 97.8% of total replies are captured within a one-week window before and after the display of community notes. We consider 10% of this window (*i.e.*, 16 hours) as the bandwidth, in which 61.3% of total replies are included, to examine the effect sizes of community notes on sentiments and emotions in our main analysis.¹¹ To further illustrate the discontinuity around the display of community notes, we plot the hourly averages of sentiments and emotions in replies within the 16-hour bandwidth before and after note display. Figs. 7.2g–7.2i demonstrate the significant discontinuities in negative sentiment, anger, and surprise at the cut-off point with stable levels at both sides (see Fig. S7.2 in Suppl. S7.1.2 for the visualizations of positive sentiment and other emotions).

11: Given the stable trends of sentiments and emotions in replies over time, we verify the robustness of RDD estimates across different bandwidths and find qualitatively identical results throughout the entire reply lifespan (see Fig. 7.5).

To shed light on the relationships between sentiments and basic emotions in our reply dataset, we examine their cross-correlations based on Pearson's r correlation coefficients. We find that anger ($r = 0.406$, $p < 0.001$), disgust ($r = 0.323$, $p < 0.001$), fear ($r = 0.089$, $p < 0.001$), and sadness ($r = 0.192$, $p < 0.001$) have significantly positive correlations with negative sentiment, while joy ($r = 0.572$, $p < 0.001$) and surprise ($r = 0.044$, $p < 0.001$) are positively correlated with positive sentiment (see Table S7.5 in Suppl. S7.1 for all cross-correlations). This suggests that, in replies to misleading posts, negative feelings are often expressed through anger, disgust, fear, and sadness. Notably, anger and disgust have the strongest correlations with negative sentiment. We further examine the cross-correlations among the six basic emotions in each reply. We find that only anger and disgust have a significantly positive correlation with each other ($r = 0.111$, $p < 0.001$). Other emotions all show negative correlations (reported in Table S7.6 in Suppl. S7.1). This suggests that replies to misleading posts often express anger and disgust together, which is a strong indicator of moral outrage [311].

7.4.2 Analysis of Sentiments and Basic Emotions (RQ1.1)

Descriptive analysis.

12: We use the t -test rather than non-parametric alternatives such as the Mann–Whitney U -test or the Kolmogorov–Smirnov test due to its advantage in reporting confidence intervals, which enhances the robustness of statistical testing for large-scale data. Although the distributions of emotions and sentiments are not strictly normal, their sample means are approximately normally distributed in our dataset. Therefore, the t -test is valid in our study based on the central limit theorem [363, 364].

We start by using t -tests to descriptively report changes in sentiments and emotions in replies after the display of community notes, compared to before note display.¹² For sentiments, positive sentiment in replies after the display of community notes (mean of 0.115) is significantly lower than before the note display (mean of 0.127; $t = 10.249$, $p < 0.001$; Fig. S7.3a). In contrast, negative sentiment in replies after the display of community notes (mean of 0.566) is significantly higher than before the note display (mean of 0.524; $t = -23.748$, $p < 0.001$; Fig. S7.3b). In terms of basic emotions, anger (mean of 0.179) and disgust (mean of 0.093) in replies after the display of community notes are significantly higher than anger (mean of 0.155; $t = -18.531$, $p < 0.001$; Fig. S7.3c) and disgust (mean of 0.090; $t = -2.717$, $p < 0.01$; Fig. S7.3d) in replies before the note display, respectively. Fear, joy, and sadness have no statistically significant changes after the display of community notes (each $p > 0.05$; see Figs. S7.3e–S7.3g). Surprise in replies after the display of community notes (mean of 0.179) is significantly lower than that in replies before the note display (mean of 0.185; $t = 5.325$, $p < 0.001$; Fig. S7.3h). Taken together, our descriptive analysis suggests that replies after note display tend to be more negative and embed higher anger and disgust, compared to replies before the display of community notes (see details in Suppl. S7.2).

Causal analysis.

We now use regression models based on RDD to estimate the causal changes in sentiments and emotions in replies to source misleading posts that are attributed to the display of community notes. The coefficient estimates for the regressions across sentiments and basic emotions are visualized in Fig. 7.3. Additionally, to interpret the effect sizes of the display of community notes on the specific sentiments and emotions in replies to misleading posts, we examine the extent to which the sentiments and basic emotions in replies change compared to their corresponding baselines before the display of community notes. The changes in sentiments and emotions in replies to misleading posts within the 16-hour window, along with the predicted effects of displaying community notes, are visualized in Fig. 7.4.

Effect of community notes display: For positive sentiment in replies, the coefficient estimate of *Displayed* is significantly negative (Fig. 7.3a: $coef. = -0.055$, $p < 0.01$; 95% CI: $[-0.089, -0.021]$). This means that the display of community notes decreases the positive sentiment in replies by 0.012 (95% CI: $[-0.020, -0.005]$). As shown in Fig. 7.4a, the baseline of positive sentiment in replies before note display is 0.125, and the positive sentiment in replies after note display is predicted to be 0.113 (95% CI: $[0.106, 0.121]$). This suggests that positive sentiment in replies decreases by 9.7% after the display of community notes compared to that before note display. In contrast, for negative sentiment in replies, the coefficient estimate of *Displayed* is significantly positive (Fig. 7.3b: $coef. = 0.114$, $p < 0.001$; 95% CI: $[0.070, 0.158]$). This means that the display of community notes increases the negative sentiment in replies by 0.038 (95% CI: $[0.023, 0.053]$). As shown in Fig. 7.4b, the baseline of negative sentiment in replies before note display is 0.521, and the negative sentiment in replies after note display is predicted to be 0.559, (95% CI:

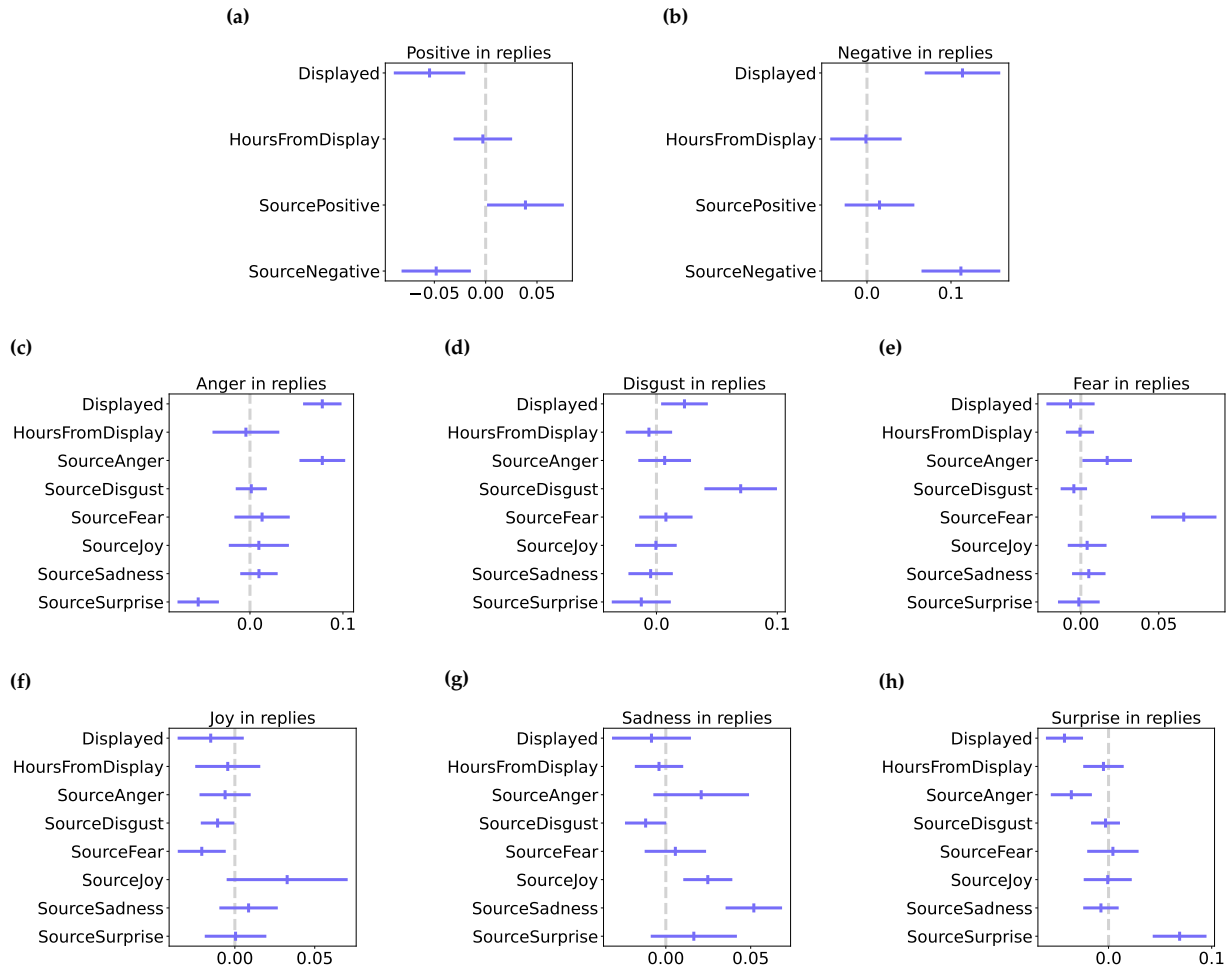


Figure 7.3: The estimated coefficients for the independent variables – *Displayed*, *HoursFromDisplay*, and source sentiments (or emotions). The independent variable *PostAge* is included during estimation but omitted in the visualization for better readability. Shown are mean values with error bars representing 95% CIs. Standard errors are clustered by source posts. The dependent variables are (a) positive sentiment in replies, (b) negative sentiment in replies, (c) anger in replies, (d) disgust in replies, (e) fear in replies, (f) joy in replies, (g) sadness in replies, and (h) surprise in replies, respectively. The full estimation results are reported in Suppl. S7.3.

[0.545, 0.574]). This suggests that negative sentiment in replies increases by 7.3% after the display of community notes compared to that before note display.

We further examine the effects of community notes display on changes in basic emotions in replies to misleading posts before and after the display of community notes. The results are as follows:

- For anger in replies, the coefficient estimate of *Displayed* is significantly positive (Fig. 7.3c: *coef.* = 0.078, $p < 0.001$; 95% CI: [0.058, 0.098]). This means that the display of community notes increases anger in replies by 0.020 (95% CI: [0.015, 0.026]). As shown in Fig. 7.4c, the baseline of anger in replies before note display is 0.154, and the predicted anger in replies after note display is 0.174 (95% CI: [0.169, 0.180]). This suggests that anger in replies increases by 13.2% after the display of community notes compared to that before note display.
- For disgust in replies, the coefficient estimate of *Displayed* is significantly positive (Fig. 7.3d: *coef.* = 0.023, $p < 0.05$; 95% CI: [0.004, 0.042]). This means that the display of community notes increases disgust in replies by 0.004 (95% CI: [0.001, 0.008]). As shown in Fig. 7.4d, the baseline of disgust in replies before note display is 0.090, and

the predicted disgust in replies after note display is 0.094 (95% CI: [0.091, 0.098]). This suggests that disgust in replies increases by 4.7% after the display of community notes compared to that before note display.

- For fear, joy, and sadness in replies, the coefficient estimates of *Displayed* are all not statistically significant (each $p > 0.05$; see Figs. 7.3e–7.3g). This means that the display of community notes has no significant effect on these three emotions in replies to misleading posts.
- For surprise in replies, the coefficient estimate of *Displayed* is significantly negative (Fig. 7.3h: *coef.* = -0.043 , $p < 0.001$; 95% CI: $[-0.060, -0.025]$). This means that the display of community notes decreases surprise in replies by 0.011 (95% CI: $[-0.015, -0.006]$). As shown in Fig. 7.4h, the baseline of surprise in replies before note display is 0.184, and the predicted surprise in replies after note display is 0.174 (95% CI: $[0.169, 0.178]$). This suggests that surprise in replies decreases by 5.7% after the display of community notes compared to that before note display.

Taken together, the display of community notes increases 7.3% more negative sentiment and triggers 13.2% more anger and 4.7% more disgust in replies to misleading posts, respectively. The effects of community notes display on sentiments and emotions in replies are concordant with the observed changes of sentiments and emotions in replies after the display of community notes (see Fig. S7.3 in Suppl. S7.2).

Effect of sentiments and emotions in source posts: We further examine the link between the original sentiments and emotions in the source misleading posts and the subsequent sentiments and emotions in the direct replies. For positive sentiment in replies, the coefficient estimate of *SourcePositive* in Fig. 7.3a is significantly positive (*coef.* = 0.039 , $p < 0.05$; 95% CI: $[0.002, 0.075]$); for negative sentiment in replies, the coefficient estimate of *SourceNegative* in Fig. 7.3b is also significantly positive (*coef.* = 0.112 , $p < 0.001$; 95% CI: $[0.066, 0.158]$). This means that the positive and negative sentiments in replies are positively linked to the original positive and negative sentiments in the source misleading posts, respectively. In terms of emotions, the coefficient estimate of *SourceAnger* in Fig. 7.3c (*coef.* = 0.078 , $p < 0.001$; 95% CI: $[0.054, 0.102]$), the coefficient estimate of *SourceDisgust* in Fig. 7.3d (*coef.* = 0.070 , $p < 0.001$; 95% CI: $[0.040, 0.099]$), the coefficient estimate of *SourceFear* in Fig. 7.3e (*coef.* = 0.066 , $p < 0.001$; 95% CI: $[0.045, 0.087]$), the coefficient estimate of *SourceSadness* in Fig. 7.3g (*coef.* = 0.052 , $p < 0.001$; 95% CI: $[0.036, 0.068]$), and the coefficient estimate of *SourceSurprise* in Fig. 7.3h (*coef.* = 0.069 , $p < 0.001$; 95% CI: $[0.043, 0.094]$) are all significantly positive. This suggests that the expressions of anger, disgust, fear, sadness, and surprise in source misleading posts have significantly positive effects on anger, disgust, fear, sadness, and surprise in the subsequent replies, respectively. However, the coefficient estimate of *SourceJoy* in Fig. 7.3f is not statistically significant (*coef.* = 0.033 , $p = 0.084$; 95% CI: $[-0.004, 0.070]$). Thus, we find no evidence that the expression of joy in replies is significantly linked to the expression of joy in the source misleading posts.

Effect of reply timing: We examine whether the sentiments and emotions in replies change over time without the display of community notes. We find that the coefficient estimates of *HoursFromDisplay* are not statistically significant across all dependent variables (each $p > 0.05$; see Figs. 7.3a–7.3h). The coefficient estimates of *PostAge* are also not statistically significant (see details in Suppl. S7.3). This means that the sentiments and emotions in replies are relatively stable over time if community notes are not displayed on the corresponding misleading posts.

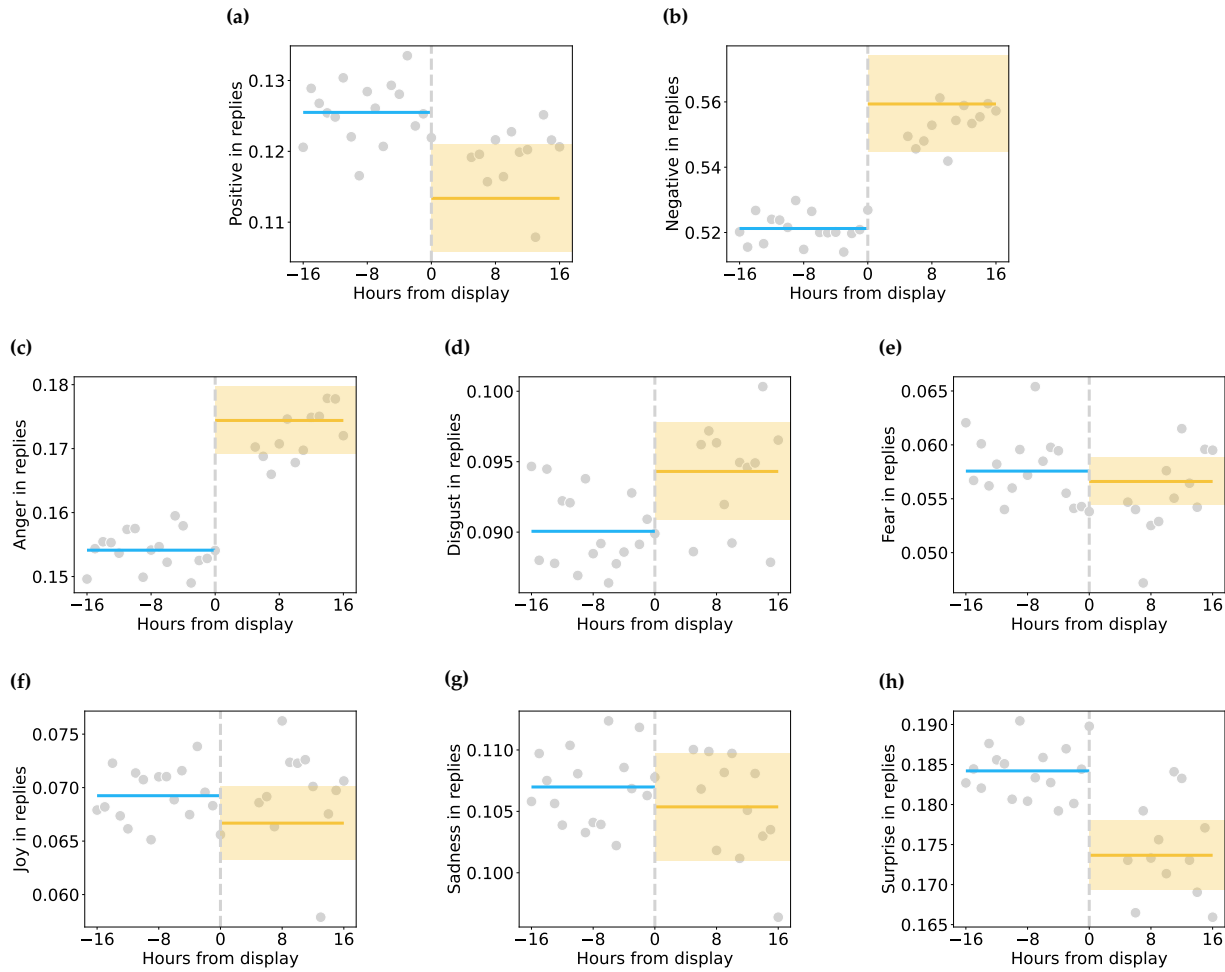


Figure 7.4: The predicted effects of the display of community notes on sentiments and emotions in replies. (a) The predicted effect on positive sentiment in replies. (b) The predicted effect on negative sentiment in replies. (c) The predicted effect on anger in replies. (d) The predicted effect on disgust in replies. (e) The predicted effect on fear in replies. (f) The predicted effect on joy in replies. (g) The predicted effect on sadness in replies. (h) The predicted effect on surprise in replies. The grey points indicate the hourly averages of sentiments or emotions between 16 hours before and 16 hours after the display of community notes. Similar to Fig. 7.2, we visually omit reply points within the initial four hours after the display of community notes for better readability. The blue lines indicate the averages of sentiments or emotions in replies over the 16 hours before note display and represent the baselines during before-display period. The yellow lines indicate the averages of sentiments or emotions over the 16 hours after the display of community notes. The average at the yellow line is the sum of the predicted effect and the corresponding baseline at the blue line in each figure. The yellow bands represent 95% CIs. The predicted effects are transformed based on the coefficient estimates of *Displayed* (see Fig. 7.3). The full estimation results are reported in Suppl. S7.3.

RDD estimates across different bandwidths: Based on the previous analysis of stable trends in sentiments and emotions within the reply lifespan (see Section 7.4.1), the RDD estimates across different bandwidths should remain robust and have no statistically significant differences. To validate this, we examine RDD estimates using three different bandwidths: 16 hours (16H), one week (1W), and the entire reply lifespan (LSP). As shown in Fig. 7.5, the RDD coefficient estimates across sentiments and emotions in replies have no statistically significant differences within the three bandwidths (see full estimation results in Suppl. S7.4). Additionally, we conduct *SUEST* tests to compare RDD models across the three bandwidths, further confirming the equivalence of the common coefficient estimates within these models (each $p > 0.05$, see Table 7.2).

Summary of findings: In sum, we find that replies are more negative and express more anger and disgust after the display of community

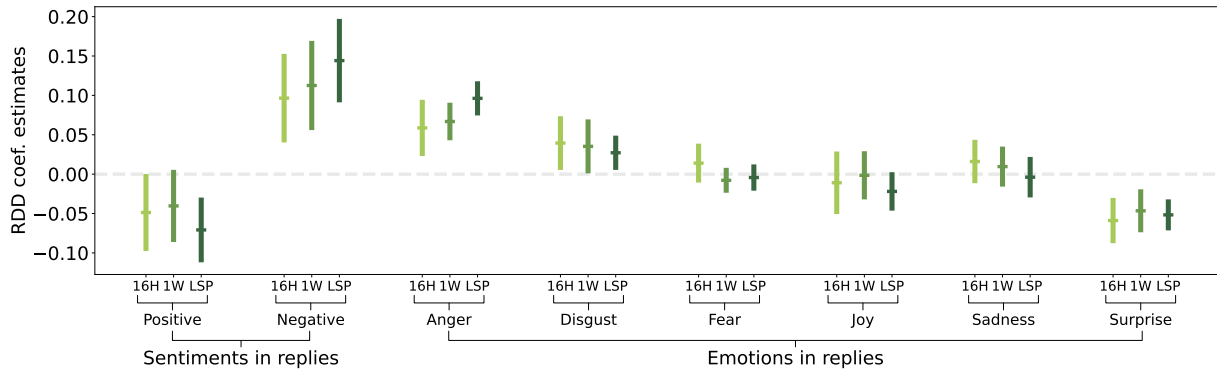


Figure 7.5: The RDD coefficient estimates (*i.e.*, *Displayed*) for sentiments and emotions in replies across three different bandwidths. The three bandwidths are 16 hours (16H), 1 week (1W), and reply lifespan (LSP), respectively. Shown are mean values with error bars representing 95% CIs. Standard errors are clustered by source posts. The full estimation results are reported in Suppl. S7.4.

Table 7.2: *SUEST* tests for the equality of common coefficients across RDD models with different bandwidths. The tests are conducted separately for each sentiment/emotion in replies.

	(1) 16H vs. LSP			(2) 1W vs. LSP		
	χ^2	p	#Variables	χ^2	p	#Variables
Sentiments in replies						
Positive	8.47	0.132	5	10.74	0.057	5
Negative	3.97	0.554	5	4.30	0.507	5
Emotions in replies						
Anger	12.87	0.168	9	15.89	0.069	9
Disgust	9.03	0.434	9	8.25	0.509	9
Fear	6.23	0.717	9	8.16	0.518	9
Joy	8.10	0.524	9	8.45	0.490	9
Sadness	13.68	0.134	9	6.70	0.668	9
Surprise	6.49	0.690	9	7.54	0.581	9

notes compared to those before the display. Specifically, the display of community notes triggers 13.2% more anger and 4.7% more disgust in replies to source misleading posts. Additionally, the sentiments and almost all emotions (except for joy) in replies are positively linked to those expressed in the source misleading posts. Furthermore, we find that the expressions of sentiments and emotions in replies are relatively stable over time if without the display of community notes.

7.4.3 Analysis Across Politic and Non-Political Misleading Posts (RQ1.2)

We conduct sensitivity analysis to examine whether the effects of community notes on sentiments and emotions in replies vary across political and non-political misleading posts. The coefficient estimates for the sensitivity analysis across political and non-political misleading posts are reported in Fig. 7.6.

We first examine the sensitivity of the effects of community notes on positive and negative sentiments. In Fig. 7.6a, the coefficient estimate of *Displayed* for positive sentiment in replies to political misleading posts is significant and negative ($coef. = -0.068$, $p < 0.01$; 95% CI:

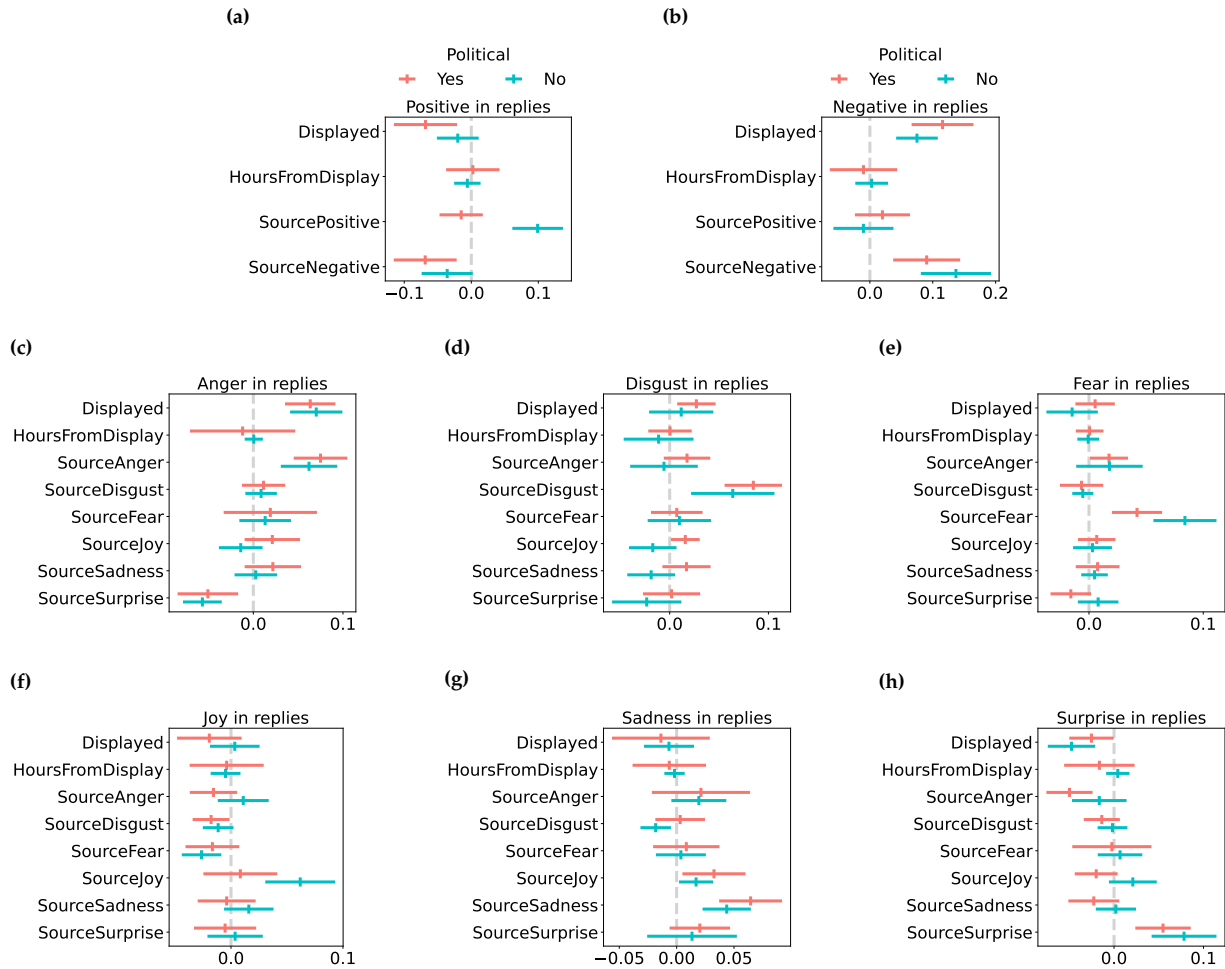


Figure 7.6: The estimated coefficients for the independent variables – *Displayed*, *HoursFromDisplay*, and source sentiments (or emotions) across political and non-political misleading posts. The independent variable *PostAge* is included during estimation but omitted in the visualization for better readability. Shown are mean values with error bars representing 95% CIs. Standard errors are clustered by source posts. The dependent variables are (a) positive sentiment in replies, (b) negative sentiment in replies, (c) anger in replies, (d) disgust in replies, (e) fear in replies, (f) joy in replies, (g) sadness in replies, and (h) surprise in replies, respectively. The full estimation results are reported in Suppl. S7.5.

$[-0.115, -0.022]$). However, the coefficient estimate of *Displayed* for positive sentiment in replies to non-political misleading posts is not statistically significant (*coef.* = -0.020 , $p = 0.197$; 95% CI: $[-0.051, 0.010]$). This means that the negative effect of community notes on positive sentiment in replies is limited to political misleading posts. Regarding the negative sentiment in replies (Fig. 7.6b), the coefficient estimates of *Displayed* are significantly positive in both political (*coef.* = 0.116 , $p < 0.001$; 95% CI: $[0.067, 0.164]$) and non-political misleading posts (*coef.* = 0.075 , $p < 0.001$; 95% CI: $[0.042, 0.107]$). This means that the effects of community notes generally apply to both political and non-political misleading posts.

Next, we examine the sensitivity of the effects of community notes on basic emotions:

- In terms of anger in replies (Fig. 7.6c), the coefficient estimates of *Displayed* are significantly positive in both political (*coef.* = 0.063 , $p < 0.001$; 95% CI: $[0.036, 0.091]$) and non-political misleading posts (*coef.* = 0.070 , $p < 0.001$; 95% CI: $[0.042, 0.099]$). This suggests that the display of community notes triggers more anger in replies to both

political and non-political misleading posts.

- In Fig. 7.6d, the coefficient estimate of *Displayed* for disgust in replies to political misleading posts is significantly positive (*coef.* = 0.027, $p < 0.01$; 95% CI: [0.008, 0.046]). However, the coefficient estimate of *Displayed* for disgust in replies to non-political misleading posts is statistically not significant (*coef.* = 0.012, $p = 0.473$; 95% CI: [-0.020, 0.044]). This indicates that the effect of community notes on increasing disgust in replies is only applicable to political misleading posts.
- As shown in Figs. 7.6e–7.6g, the coefficient estimates of *Displayed* for fear, joy, and sadness in replies to both political and non-political misleading posts are consistently not significant (each $p > 0.05$). This means that the display of community notes has no significant effects on fear, joy, and sadness in replies to either political or non-political misleading posts.
- In Fig. 7.6h, the coefficient estimates of *Displayed* for surprise in replies to both political (*coef.* = -0.025, $p < 0.05$; 95% CI: [-0.050, -0.001]) and non-political misleading posts (*coef.* = -0.048, $p < 0.001$; 95% CI: [-0.074, -0.022]) are significantly negative. This means that the display of community notes generally leads to a decrease of surprise in replies to both political and non-political misleading posts.

Additionally, we examine the associations between the sentiments and emotions in replies and the source sentiments and emotions across political and non-political misleading posts. Fig. 7.6a) shows that the coefficient estimate of *SourcePositive* is not statistically significant for positive sentiment in replies to political misleading posts (*coef.* = -0.015, $p = 0.350$; 95% CI: [-0.047, 0.017]), while it is significantly positive for positive sentiment in replies to non-political misleading posts (*coef.* = 0.099, $p < 0.001$; 95% CI: [0.062, 0.136]). Similarly, in Fig. 7.6f, the coefficient estimate of *SourceJoy* is statistically not significant for joy in replies to political misleading posts (*coef.* = 0.008, $p = 0.614$; 95% CI: [-0.024, 0.041]), while it is significantly positive for joy in replies to non-political misleading posts (*coef.* = 0.062, $p < 0.001$; 95% CI: [0.031, 0.092]). This suggests that the positive association between source positive sentiment (joy) in original posts and positive sentiment (joy) in replies disappears for political misleading posts. In contrast, for negative sentiment, anger, disgust, fear, sadness, and surprise in replies, there are consistently positive associations with their corresponding emotions in the source misleading posts, regardless of whether the posts are political or non-political (each $p < 0.05$, see Figs. 7.6b–7.6e and Figs. 7.6g–7.6h).

Finally, we find that the coefficient estimates of *HoursFromDisplay* for all sentiment/emotions in replies to either political or non-political misleading posts are consistently not significant (each $p > 0.05$; see Fig. 7.6). This suggests that the stability of sentiments and emotions over time remains robust, irrespective of whether the misleading posts are political or non-political.

Summary of findings: Taken together, we observe two aspects of our findings that are specific to political misleading posts: (i) The effects of community notes on the decrease of positive sentiment and the increase of disgust in replies only apply to political misleading posts; (ii) The transfer of positive sentiment, particularly joy, from source posts to replies is disrupted in discussions on political topics.

7.4.4 Analysis of Moral Outrage (RQ2.1 & RQ2.2)

Given that moral outrage is typically expressed through anger and disgust, we examine whether the display of community notes triggers the moral outrage of users in their replies to misleading posts. Analogous to previous research [310], we define the dependent variable moral outrage as $MoralOutrage = Anger \times Disgust$. We first examine the changes of moral outrage in replies to all misleading posts. As shown in Fig. 7.7a, the coefficient estimate of *Displayed* is significantly positive ($coef. = 0.007$, $p < 0.001$; 95% CI: [0.005, 0.008]). This means that the display of community notes increases moral outrage in replies to misleading posts by 0.003 (95% CI: [0.002, 0.004]). Additionally, using the baseline of moral outrage during a period of 16 hours before note display (mean of 0.019, Fig. 7.7b), we find that the display of community notes triggers 16% more moral outrage in replies and increases it to a value of 0.023 (95% CI: [0.022, 0.023]).

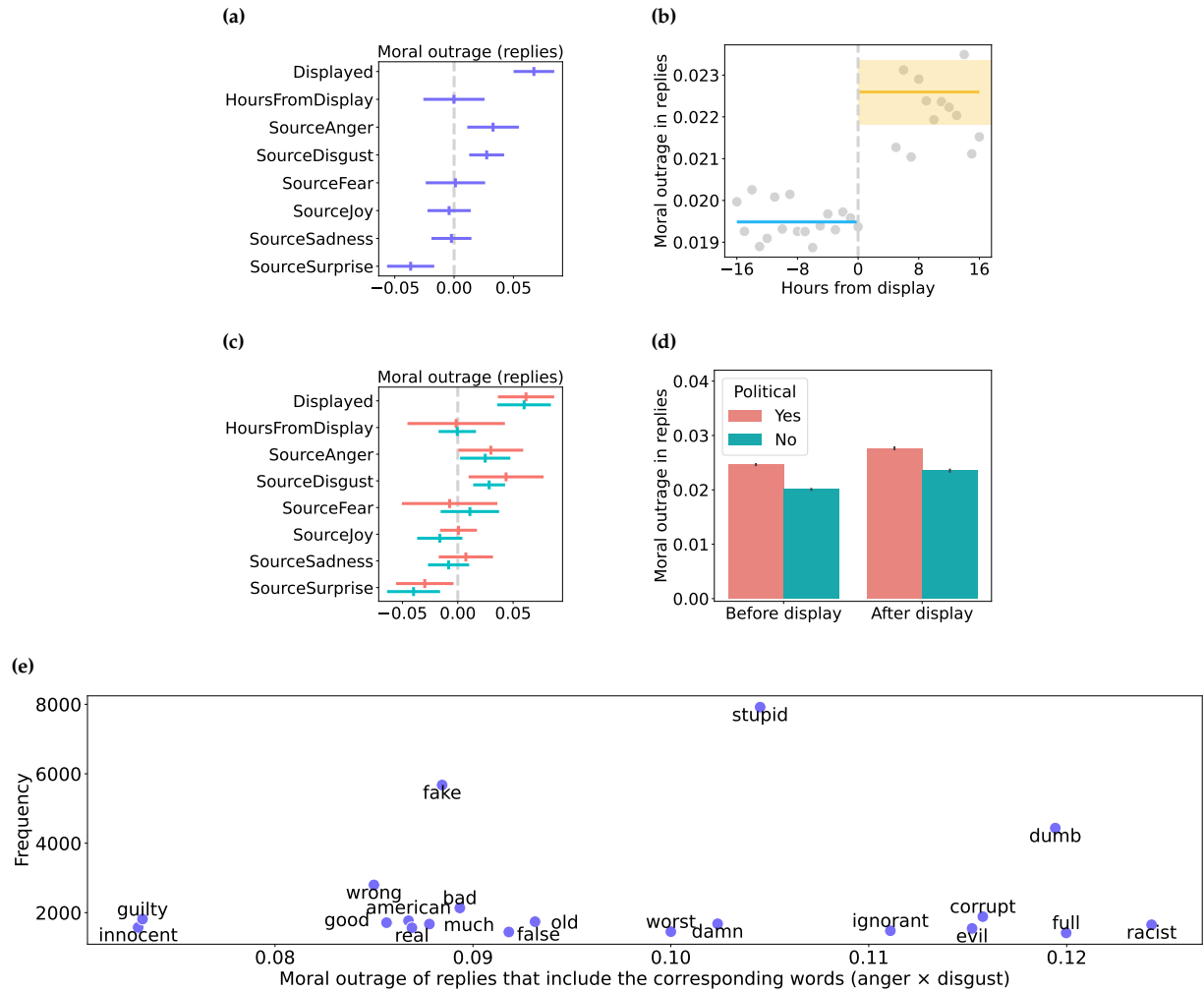


Figure 7.7: Moral outrage in replies to misleading posts increases after the display of community notes. (a) The estimated coefficients for the independent variables – *Displayed*, *HoursFromDisplay*, and source emotions. (b) The predicted effect of the display of community notes on moral outrage in replies to misleading posts. The error band represents 95% CIs. (c) The estimated coefficients for the independent variables – *Displayed*, *HoursFromDisplay* across political and non-political misleading posts. (d) The averages of moral outrage in replies to political and non-political misleading posts before and after the display of community notes. The error bars represent 95% CIs. (e) The twenty most frequent adjectives within replies with high moral outrage. For (a) and (c), the independent variable *PostAge* is included during estimation but omitted in the visualization for better readability. Shown are mean values with error bars representing 95% CIs. Standard errors are clustered by source posts. The full estimation results are reported in Suppl. S7.6.

Robustness with alternative classifiers: Previous research developed a dedicated classifier to predict moral outrage, namely, the Digital Outrage Classifier (DOC) [312]. As a robustness check, we repeat our analysis with this classifier and find that the coefficient estimate of *Displayed* remains significantly positive ($coef. = 0.067, p < 0.001$; 95% CI: [0.041, 0.092]). A *SUEST* test indicates no statistically significant difference in the effect of *Displayed* between the operationalization of moral outrage in our main analysis and the operationalization of moral outrage via the DOC classifier ($\chi^2 = 0.00, p = 0.957$). Additionally, anger and disgust are central components of other-condemning moral emotions [81]. Using a dedicated moral emotion classifier developed by prior research [365], we repeat our analysis and find that the effect of *Displayed* on other-condemning moral emotions is also significantly positive ($coef. = 0.132, p < 0.001$; 95% CI: [0.103, 0.161]). Altogether, these results consistently support that community notes trigger moral outrage beyond the isolated effects on anger and disgust (see details in Suppl. S7.6).

Analysis across political and non-political misleading posts: We further examine the effects of community notes display on moral outrage in replies to political and non-political misleading posts separately. As shown in Fig. 7.7c, the coefficient estimates of *Displayed* are significantly positive within both political ($coef. = 0.062, p < 0.001$; 95% CI: [0.037, 0.087]) and non-political misleading posts ($coef. = 0.060, p < 0.001$; 95% CI: [0.036, 0.084]). This indicates that the effects of community notes in triggering moral outrage are robust across political and non-political misleading posts. Additionally, we analyze the average levels of moral outrage in replies to political and non-political misleading posts before and after the display of community notes (Fig. 7.7d). Before note display, the averages of moral outrage in replies to political and non-political misleading posts are 0.025 and 0.020, respectively ($t = 63.709, p < 0.001$). After note display, these averages increase to 0.028 for political posts and 0.024 for non-political posts ($t = 31.042, p < 0.001$). This suggests that the moral outrage in replies to political misleading posts is significantly higher than that in replies to non-political misleading posts, regardless of the display of community notes. However, the averages of moral outrage in the source posts between political (mean of 0.010) and non-political ones (mean of 0.009; $t = 0.575, p = 0.566$) have no significant difference. This indicates that the heightened moral outrage in replies to political posts, compared to non-political posts, is not simply a transfer from the source posts' moral outrage.

Validation of targets of moral outrage: We conduct multiple tests to validate that the measured moral outrage in replies is primarily directed at the misleading posts or their authors (rather than reflecting emotional reactions from users whose "side" has been labeled as misleading).

First, we consider two scenarios where replies may not be directed toward original posts: mentions of other users (not authors) in replies and mentions of community notes in replies. If the level of moral outrage in these two scenarios is lower than in replies directed solely toward original posts, this would suggest that the moral outrage in replies is primarily focused on misleading content. We find that the ratio of mentioning other users in replies after note display (10.4%) is lower than that in replies before note display (12.3%). Meanwhile, after note display, moral outrage in replies that do not mention others (mean of 0.028) is 180.0% more than that in replies that mention others (mean of 0.010; $t = -82.018, p < 0.001$). Additionally, after the display of community notes, replies may discuss community notes rather than the original misleading posts. Given this, we check how many replies

mention community notes after their display through the keywords: “community note” and “CN.” We find that, out of 533,495 replies created after note display, only 2986 (0.6%) replies mention community notes. Moreover, moral outrage in replies that do not mention community notes (mean of 0.026) is 73.3% higher than the moral outrage in replies that mention community notes (mean of 0.015; $t = -11.791$, $p < 0.001$). Taken together, these validations suggest that the majority of replies, especially those with high moral outrage, are directed toward source misleading posts.

Second, we randomly select 100 misleading posts with displayed community notes and manually evaluate the reply with the highest outrage score after the display of community note for each selected post. Specifically, two authors of this study independently reviewed these 100 replies, arriving at identical results: all 100 replies are consistently directed toward the content of the misleading posts or their authors.

Third, we analyze the most frequently used words in replies that express moral outrage. We extract all words from replies posted after note display, focusing on those with a higher level of moral outrage (*i.e.*, above the mean value). The most frequent word in these replies is “you,” often indicating a direct reference to the post author. Subsequently, we remove stop words and use Spacy, a Python library for advanced natural language processing, to extract the twenty most frequent adjectives. Then, we calculate the average level of moral outrage for each word based on the replies that include it. Fig. 7.7e shows that “stupid,” “fake,” and “dumb” are the most frequent adjectives in replies expressing moral outrage, with mean values of 0.105, 0.088, and 0.119, respectively. Additionally, words associated with higher levels of moral outrage (greater than 0.100) tend to be more critical (*e.g.*, racist, dumb, and corrupt), while words associated with lower levels of moral outrage (below 0.100) are more related to the veracity of the content (*e.g.*, wrong, fake, and false). This suggests that higher levels of moral outrage are associated with the use of more critical or aggressive words.

Summary of findings: Altogether, our analysis suggests that the display of community notes triggers moral outrage in replies to the corresponding misleading posts. Additionally, political misleading posts tend to receive more moral outrage compared to non-political misleading posts. However, the effects of community notes on moral outrage in replies are robust across political and non-political misleading posts.

7.4.5 Robustness Checks

We conduct a wide range of additional robustness checks to ensure the reliability of our findings:

- **Multicollinearity check:** We check for possible multicollinearity issues using Variance Inflation Factors (VIFs). The VIFs are all close to one; and, thus well below the critical threshold of four (see details in Suppl. S7.7.1).
- **Analysis with lagged-dependent variables:** To address the possibility that users replying to original posts may be influenced by sentiments and emotions in previous direct replies, we conduct an auto-correlation test and repeat our analysis with lagged-dependent variables. The results remain robust, consistently supporting our main findings (see details in Suppl. S7.7.2).

- **Shorter bandwidth and placebo test:** To further ensure that the increase in moral outrage in replies to misleading posts is specific to the display of community notes, we repeat our analysis within a shorter bandwidth, *i.e.*, 2 hours, around the display of community notes and conduct a placebo test at a different cut-off point before note display for comparison. We find that all results are robust and consistent with our main analysis (see details in Suppl. S7.7.3).
- **Analysis without reply restriction:** We repeat the analysis without the restriction of 5 replies before and after note display. The results remain robust and consistent with our main findings (see details in Suppl. S7.7.4).
- **Month-year fixed effects:** We incorporate month-year fixed effects into our regression models and repeat our analysis. The results remain robust and consistent with our main findings (see details in Suppl. S7.7.5).
- **Validation with dictionary-based approaches:** We repeat our analysis using sentiment and emotion lexicons. Specifically, we use the VADER lexicon for sentiments and the NRC lexicon for emotions [163, 366]. The results remain robust across machine learning models and lexicons (see details in Suppl. S7.7.6).
- **Role of sentiments and emotions in community notes:** To explore whether sentiments and emotions in community notes can potentially moderate the effects of community notes display, we incorporate sentiments and emotions in community notes and their interactions with *Displayed* into our regression models. The coefficient estimates of *Displayed* across the dependent variables remain robust. Additionally, the majority of the coefficient estimates for the interactions between sentiments (emotions) in community notes and *Displayed* are not statistically significant. This suggests that users' emotional reactions are not significantly affected by the sentiments or emotions in community notes (see details in Suppl. S7.7.7).
- **Robustness across helpfulness scores:** Misleading posts that garner significant public concern may be more likely to be quoted by influential accounts on $\mathbb{X}$, thereby reaching a broader audience. This increased visibility may introduce confounding factors related to the helpfulness scores of community notes. For example, counter-partisan accounts can encourage other users to rate a community note as helpful, thereby increasing the helpfulness score of the note [50]. To account for such potential confounding factors, we incorporate *NoteScore* and its interaction with *Displayed* into our regression models. We find that the effect of community notes is robust across different helpfulness scores (see details in Suppl. S7.7.8).
- **Effect of community notes over the program's development:** Given that the Community Notes program continuously evolves, we incorporate *MFRO* (*i.e.*, months from the roll-out of Community Notes program) and its interaction with *Displayed* into regression models to investigate how the effect of community notes changes over time from the roll-out of the program. We find that the effect of community notes on our dependent variables remains stable (see details in Suppl. S7.7.9).

7.4.6 Summary of Main Findings

In summary, we conduct a comprehensive *causal* analysis on the role of community notes in triggering sentiments, emotions, and moral outrage in replies to fact-checked social media posts. Furthermore, we extend

our analysis by conducting sensitivity analysis across political and non-political misleading posts. Our main findings are as follows:

- ▶ The display of community notes triggers 7.3% more negativity, 13.2% more anger, and 4.7% more disgust in replies to misleading posts (RQ1.1).
- ▶ The effects of community notes on negativity and anger are robust across political and non-political misleading posts, while the effect on disgust only applies to political misleading posts (RQ1.2).
- ▶ The display of community notes triggers 16% more moral outrage (RQ2.1), and this is robust across political and non-political misleading posts (RQ2.2).

In addition to our main findings, we find the evidence of emotion transfers and time-independence for sentiments and emotions in online social networks:

- ▶ The sentiments and emotions in replies to misleading posts are positively correlated with the sentiments and emotions of the corresponding source posts (with very few exceptions).
- ▶ The sentiments and emotions in replies are stable over time if the posts are not subject to fact-checking (*i.e.*, do not receive a displayed community note).

7.5 Discussion

7.5.1 Relevance

The spread of online misinformation has become a significant challenge that social media platforms have to tackle. However, traditional (expert-based) fact-checking approaches have limitations regarding their scalability and face trust issues among the user base. As a remedy, crowdsourced fact-checking represents a promising approach to identify misinformation at scale and increase users' trust in fact-checks. As the first large-scale attempt of crowdsourced fact-checking approach on a major social media platform, the "Community Notes" program has been shown to be effective in producing trustworthy community fact-checks and reducing engagement with misleading posts on X [56, 76]. It provides a reference for other social media platforms to implement similar crowdsourced fact-checking approaches (*e.g.*, YouTube [367]). However, to improve and extend the "Community Notes" program safely, it is important to understand how community notes affect the discussion environment after they are displayed on misleading posts, *i.e.*, how users' emotional expressions change after they are informed about the falsehood. To this end, our study causally examines the changes in emotional expressions in replies to misleading posts before and after the display of community notes. As detailed in the next sections, our findings provide important insights into the effects of (community-based) fact-checking on the discussion environment on social media and may help to improve the design of crowdsourced fact-checking platforms.

7.5.2 Interpretations

Community notes trigger negativity, anger, and disgust in replies to misleading posts. We find that the display of community notes leads users to post more negative replies, which is primarily due to

heightened expressions of anger and disgust. These observations align with previous survey and lab studies in the field of psychology that people often report anger and disgust in response to moral violations of community, autonomy, and divinity ethics [368, 369]. In the context of our study, the display of community notes signals to users that the posts they are exposed to spread misinformation, which may thereby trigger their perception of a moral violation. Anger and disgust, though both negative, differ in their aggressive tendencies when responding to moral violations. Anger is typically associated with high-cost, direct aggression (*e.g.*, physical violence or confrontational responses), whereas disgust is typically associated with less costly indirect aggression (*e.g.*, gossip and social exclusion). Additionally, previous research suggests that people tend to express more anger than disgust when moral violations directly affect themselves, while they express more disgust than anger when moral violations target others [368]. Quantitatively, our findings reveal that the display of community notes increases anger in replies to misleading posts by 13.2%, compared to a 4.7% increase in disgust. Thus, users are more likely to express anger than disgust, which may indicate that users perceive the misleading content to primarily impact their own interests (rather than others'). Notably, our sensitivity analysis further reveals that the effect of community notes in increasing disgust is statistically significant only in replies to political misleading posts. This could partially explain why the display of community notes is less effective in reducing the spread of political misleading posts, compared to non-political ones, which has been observed in a previous study [76].

Community notes trigger moral outrage in replies to misleading posts.

When anger and disgust are mixed, they generate “moral outrage,” a more intense emotional reaction to perceived moral transgressions than either anger or disgust alone [310–312]. Our findings show that the display of community notes increases moral outrage by 16% in replies to misleading posts, a greater increase than those observed for anger (13.2%) and disgust (4.7%) individually. This empirical evidence underscores the significant role of moral outrage in users' responses to the moral transgression of spreading misleading posts. Specifically, our findings suggest that displaying community notes on misleading posts is a strong stimulus that calls attention to the moral norm violation, thereby motivating expressions of moral outrage. Notably, research suggests that if users experience moral violations personally without such external stimuli, their intentions to express moral outrage might decrease [310]. This phenomenon might be explained by social conformity that users are inclined to conform to the majority's opinions on the quality of online posts [370, 371]. The display of community notes signals consensus among diverse users, thereby potentially encouraging exposed users' conformity and subsequent correcting actions.

The expression of moral outrage to spreading misinformation is a double-edged sword that can both mitigate the harm of online misinformation and exacerbate the polarization of online communities. On the one hand, moral outrage can strengthen debunking effects on exposed users and pressure authors to correct or delete their misleading posts [76, 372]. Previous research has shown that the display of community notes significantly increases the likelihood of deleting misleading posts by their authors [76]. Moreover, those who attempt to spread misinformation can expect strong negative reactions from the user base once their content is debunked, which may help to discipline users and encourage more responsible behavior on social media. On the other hand, social media platforms also need to maintain a healthy, inclusive, and balanced online

environment. However, online moral outrage can deepen social divides and fuel political polarization [33, 310, 312]. Additionally, the corrections and moral judgments from peer users may backfire and make authors of misleading posts feel defensive, promoting their subsequent shares of low-quality, partisan, and toxic content [373]. In this light, while moral outrage can be a powerful tool for combating misinformation, it also risks escalating conflicts and entrenching divisions within online communities. This risk may partially explain why community notes lead to smaller reductions in replies compared to reposts [304]. The potential negative effects of moral outrage triggered by community notes require a further rigorous examination to find a trade-off between mitigating the spread of misinformation and maintaining a positive discussion atmosphere.

Emotion transfers and stable emotional responses over time. Our study provides strong empirical evidence supporting the idea that connected users on social media platforms express similar sentiments/emotions. Specifically, we find that all sentiments and emotions in misleading posts, especially non-political ones, are effectively transferred to subsequent sentiments and emotions in replies. This emotion transfer phenomenon can be explained primarily through two mechanisms: emotion contagion and emotion alignment (*i.e.*, homophily) among connected users. Here, emotion contagion refers to the spreading effect of emotions from one user to another [356, 358], and emotion homophily indicates that like-minded users tend to flock together, potentially forming echo chambers that dominate online interactions and reinforce biases and attitude polarization [29]. Our findings further reveal that the intensity of sentiments and emotions in replies remains stable over time, showing no signs of fading. Moreover, misleading posts fact-checked via community notes are predominantly characterized by negativity. The persistence of emotion transfers may further exacerbate these dynamics [10, 335, 336].

7.5.3 Implications for Further Design of Community-Based Fact-Checking

Community-based fact-checking is a promising approach and can effectively reduce engagement with misinformation on social media [46, 55, 56, 76, 329]. However, our study finds that communicating falsehoods to users via community fact-checks can also trigger moral outrage in response to the perceived moral violation of spread misinformation. This heightened moral outrage can have dual impacts. On the one hand, it may strengthen the effectiveness of community-based fact-checking by pressuring (disciplining) users to reconsider their actions. On the other hand, however, it may also backfire by increasing tension and polarization on social media [373].

The dual role of moral outrage highlights the need to minimize its negative consequences without undermining its positive effects in the design and implementation of community-based fact-checking to maintain health and constructive dialogue on social media. Previous research suggests that temporarily deactivating social media accounts can reduce polarization on policy issues [105]. Additionally, a “cooldown” period for highly active and polarized threads has been proposed to ensure a deliberately interactive environment [374]. Given this, social media platforms might consider implementing a *temporary freezing period* for misleading posts once they receive helpful community fact-checks. During this period – *e.g.*, one hour – the misleading posts would be invisible to others, while the authors are notified, giving them time to delete or

correct the misleading content. If they do not respond within the given timeframe, the community fact-checks with associated misleading posts would then be publicly displayed on the platform. Similarly, platforms may consider *temporarily limiting the ability to comment* on posts flagged by community fact-checks. This could take the form of a complete ban on posting responses during this period or a personal delay, where written comments are visible to others only after a certain period. In general, platforms may need to emphasize *community guidelines* in heated/morally charged discussions (*e.g.*, labels/brief prompts that appear when a user is about to post a highly emotional response; or additional monitoring by moderation teams) [223].

Importantly, however, any such design changes must be carefully evaluated through further research to assess their potential benefits and drawbacks. A core motivation behind community-based fact-checking is to reduce censorship on social media platforms and empower users with greater autonomy to improve trust in the fact-checking process [56]. Therefore, platforms must strike a delicate balance between fostering constructive discourse and preserving user autonomy to sustain trust and engagement in community-based fact-checking.

7.5.4 Limitations and Future Research

Our study has several limitations that could open up potential avenues for future research. First, our study is limited to a 4-month period since the roll-out of the “Community Notes” program. As the program is continually evolving, it would be valuable to examine the effects of community notes on sentiments, emotions, and moral outrage over a longer observation timeframe.¹³ Furthermore, while we conduct extensive robustness checks to ensure the reliability of our findings, it is challenging for a quasi-experimental study to fully isolate the effect of community notes from other potential concurrent misinformation interventions or external factors. Therefore, future research is encouraged to validate our findings within controlled experimental designs.

Second, on the date of data collection, the “Community Notes” program was primarily well-established in the U.S., and the fact-checked posts were mainly in English. Consequently, we only consider English posts in our study. In the future, a multilingual and cross-cultural evaluation would be beneficial to get a more comprehensive understanding of how community notes impact users’ emotional reactions across languages and cultures. Additionally, our sensitivity analysis is limited to general political and non-political misleading posts. Future research could expand this scope by examining more fine-grained topics (*e.g.*, health, business & economics, celebrities, *etc.*).

Third, our measurement of moral outrage cannot definitively ascertain whether it is directed toward misleading posts or their authors. To validate our approach, we analyze two special cases: mentions of other users and mentions of community notes in replies. We find that the moral outrage in replies that mention other users or community notes is significantly lower than the moral outrage in replies that solely react to the misleading posts and do not mention other users or community notes. Additionally, replies with high moral outrage frequently use aggressive words (*e.g.*, “stupid,” “dumb”). All of these analyses support the notion that community notes trigger moral outrage and critical remarks to misleading posts or their authors. Future research could further validate

13: Notably, we collected our dataset via X’s Academic Research API before it was deprecated. Currently, it is costly to collect such a high volume of data from X Pro API.

our findings through more advanced subject-aware emotion detection models.

Finally, it has been found that the display of community notes increases the probability of post deletion by authors [76]. A possible contributing factor may be that expressions of moral outrage pressure authors to delete their misleading posts. This notion could be validated through the examination of moral outrage in replies to deleted and non-deleted posts. However, deleted misleading posts and corresponding replies cannot be collected through X's API. As a remedy, future research could conduct lab experiments to causally examine whether expressions of moral outrage increase authors' intentions to delete their misleading posts. Additionally, it would be interesting to further examine how the authors of misleading posts adapt their future behaviors on social media.

7.6 Conclusion

Community-based fact-checking is a promising approach to reduce engagement with misinformation on social media. Yet, an understanding of how community fact-checks affect users' emotional perception of misleading posts and their authors was missing. In this study, we analyze a large-scale panel dataset of replies to misleading posts that have been fact-checked via community notes. Our findings provide strong causal evidence that community fact-checks trigger moral outrage in replies to misleading posts. This suggests that social media users perceive spreading misinformation as moral transgressions, and that spreaders of misinformation should expect negative responses once their content is debunked. Our study highlights the need for further examination of both the positive and negative consequences of such emotional backlash.

PART III: RESILIENCE AND PARTICIPATION IN COMMUNITY-BASED FACT-CHECKING

Abstract: Community-based fact-checking systems, such as Community Notes on X (formerly Twitter), aim to mitigate online misinformation by surfacing annotations judged helpful by contributors with diverse viewpoints. While prior work has shown that the platform’s bridging-based algorithm effectively selects helpful notes at the time of display, little is known about how evaluations change after notes become visible. Using a large-scale dataset of 437,396 community notes and 35 million ratings from over 580,000 contributors, we examine the stability of helpful notes and the rating dynamics that follow their initial display. We find that 30.2% of displayed notes later lose their helpful status and disappear. Using interrupted time series models, we further show that note display triggers a sharp increase in rating volume and a significant shift in rating leaning, but these effects differ across rater groups. Contributors with viewpoints similar to note authors tend to increase supportive ratings, while dissimilar contributors increase negative ratings, producing systematic post-display polarization. Counterfactual analyses suggest that this post-display polarization, particularly from dissimilar raters, plays a substantial role in note disappearance. These findings highlight the vulnerability of consensus-based fact-checking systems to polarized rating behavior and suggest pathways for improving their resilience.

8.1	Introduction	153
8.2	Background and Related Work	155
8.3	Data and Methods . . .	157
8.3.1	Dataset	157
8.3.2	Reproduction of Note Selection Algorithm . .	157
8.3.3	Matrix Factorization for Rater Similarity	158
8.3.4	Estimation Models . . .	159
8.4	Results	160
8.4.1	Note Disappearance (RQ1 & RQ2)	160
8.4.2	Changes in Rating Count and Leaning (RQ3) . . .	161
8.4.3	Similar vs. Dissimilar Raters (RQ4)	162
8.4.4	Counterfactual Analysis	164
8.5	Discussion	165

Publication

Y. Chuai, G. Lenzini, and N. Pröllochs, ‘Consensus stability of community notes on X,’ in *Proceedings of the ACM on Web Conference 2026*, 2026, pp. 8885–8896.

8.1 Introduction

The proliferation of misinformation on social media platforms poses significant risks to democratic governance, public trust, and social cohesion [1, 10, 33]. In response, social media providers face mounting pressure to combat the spread of misinformation on their platforms [263]. An increasingly adopted strategy is to leverage the “wisdom of crowds,” whereby users collaboratively identify misinformation and contribute context annotations for correction through community-based (crowd-sourced) fact-checking [46]. X’s Community Notes program has emerged as a flagship implementation of this approach and has since influenced similar initiatives on other platforms, including YouTube, Facebook, and TikTok [79].

Participants in the Community Notes program (*i.e.*, *contributors*) can annotate potentially misleading posts by proposing *community notes*, which are then evaluated by other contributors. However, unlike expert fact-checkers who provide rigorous and authoritative assessments, crowd contributors vary in expertise, consistency, and impartiality across both note writing and rating processes [50, 54, 68, 69]. This creates a central challenge for community-based fact-checking: reliably surfacing accurate and broadly helpful annotations while preventing the display of low-quality, biased, or strategically manipulated notes [45]. To address this

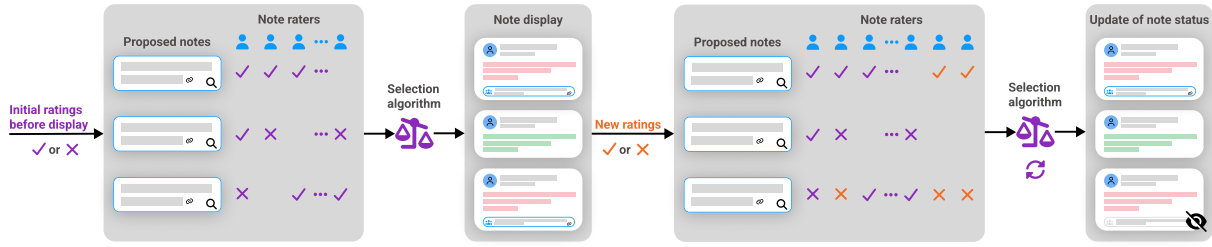


Figure 8.1: Workflow of the note selection algorithm. Community notes are first proposed and rated by contributors, and those deemed helpful are then publicly displayed to all users. The algorithm then continues to update note status as additional ratings are received, meaning displayed notes can later lose or regain their helpful status.

challenge, $\mathbb{X}$ employs a bridging-based algorithm designed to identify points of agreement across politically diverse groups and promote notes that are both factually informative and widely perceived as helpful [67]. Prior work has shown that this algorithm can effectively select accurate notes that receive broad cross-perspective supports [67, 70].

While the effectiveness of the note selection algorithm in surfacing helpful notes on $\mathbb{X}$ has been well documented, little is known about what happens *after* a note becomes visible. As illustrated in Fig. 8.1, the algorithm continuously re-evaluates community notes by incorporating new ratings (*i.e.*, upvotes and downvotes) on a rolling basis. As new ratings arrive, a note deemed helpful at one moment may later lose visibility; and, even factually accurate notes can lose consensus and disappear from view [45, 50]. This dynamic process raises the possibility that *visibility itself may reshape subsequent crowd behavior*: some contributors may reinforce a note with additional support, while others may respond with disagreement, partisan backlash, or strategic downvoting. This is further complicated by potential manipulation attempts or coordinated attacks aimed at influencing note status. Together, the dynamic rating system raise an important question about the resilience of community-based fact-checking systems: do notes that initially achieve broad consensus remain stable over time, or do post-display rating dynamics (whether organic or adversarial) erode consensus? Addressing this question is the goal of our study.

Research questions. In this study, we examine the consensus stability of community notes to assess the resilience of the note selection algorithm on $\mathbb{X}$'s Community Notes platform. Using a large-scale dataset of 437,396 Community Notes and more than 35 million ratings from over 583,000 contributors, we address four main research questions (RQs):

- ▶ **RQ1:** How often do community notes that are initially rated as helpful later lose their helpful status and disappear?
- ▶ **RQ2:** How is note disappearance related to characteristics of the underlying posts and their authors?
- ▶ **RQ3:** How do rating volume and the balance of upvotes vs. downvotes change after a note becomes visible?
- ▶ **RQ4:** Do raters with similar vs. dissimilar viewpoints evaluate the same note differently once it becomes visible?

Methodology. To address our RQs 1–2, we employ a logistic regression model to examine the likelihood of note disappearance. For RQs 3–4, we employ a design of interrupted time series to analyze the changes in the *rating count* and *rating leaning* after the initial display of community notes across different rater groups. The rating count indicates the total number of ratings submitted to a note, while the rating leaning indicates the difference between upvotes and downvotes during a certain period.

Contributions. Our analysis contributes four main findings: (i) We demonstrate that 30.2% of displayed notes later lose their helpful status and disappear. (ii) Note disappearance is systematically associated with characteristics of the source post: notes attached to posts about health or political topics, as well as posts authored by high-influence users, are significantly more likely to disappear. We also observe a political asymmetry such that notes displayed on posts from left-leaning authors are more likely to disappear than those on right-leaning authors' posts. (iii) Using interrupted time series models, we further show that note display triggers a sharp increase in rating volume and a significant shift in rating leaning. (iv) However, these shifts differ across rater groups. Contributors with viewpoints similar to note authors tend to increasingly provide supportive ratings, while dissimilar contributors increasingly provide negative ratings, producing systematic post-display polarization. Counterfactual analyses confirm that this post-display polarization, particularly from dissimilar raters, substantially contributes to note disappearance. Altogether, our findings highlight the vulnerability of consensus-based fact-checking systems to polarized rating behavior and suggest pathways for improving their resilience.

8.2 Background and Related Work

Approaches to fact-checking: Effectively identifying and mitigating the spread of online misinformation remains a persistent challenge [1]. Traditional expert-based fact-checking approaches, though capable of producing rigorous and authoritative assessments, face fundamental limitations in scalability and timeliness, rendering them insufficient for the volume and velocity of online misinformation [46, 47, 54, 74]. To overcome these constraints, scholars and practitioners have increasingly turned to community-based (or crowdsourced) fact-checking, which leverages the crowd to collaboratively identify and contextualize misleading content [46, 52, 53, 67, 281]. A growing body of research has examined expert-based (*e.g.*, PolitiFact) and community-based (*e.g.*, X's Community Notes) fact-checking systems from multiple perspectives to examine fact-check generation and evaluation [375].

Identification of misleading content: Current fact-checking efforts, whether conducted by professional experts or by crowds, remain inherently human-driven. Because someone must first decide which content warrants fact-checking, the selection process is vulnerable to subjective judgment and potential bias [69, 73, 74, 152]. Research shows that fact-checking activity tends to surge during major societal and political events such as the COVID-19 pandemic or presidential elections [73, 152]. Partisan asymmetries in fact-checking selection have been observed in third-party organizations such as PolitiFact and Snopes. For example, PolitiFact rates Republican statements as false at roughly three times the rate of Democratic statements [159]. Additionally, fact-checked false statements from Republicans are disproportionately those that reference Democrats, oftentimes with the expression of out-group animosity [73, 105].

Similar partisan asymmetries also appear in community-based fact-checking systems [376, 377]. On X's Community Notes platform, contributors are more likely to flag posts from Republicans than those from Democrats [378], and they tend to write community notes on posts from counter-partisans [50]. A recent study finds that a new feature on X, *i.e.*,

“Request Community Note,” introduces further partisan divergence: contributors more often annotate posts from Republicans, whereas *requestors* who disproportionately submit requests targeting Democratic posts [79]. Collectively, these findings suggest that both expert and crowd-based fact-checking processes exhibit partisan asymmetries in fact-checking selection. However, whether these asymmetries reflect inherent selection bias or the broader misinformation landscape remains an open question [73]. Additionally, beyond partisanship, comparative analyses indicate that community-based and professional fact-checking systems, such as Community Notes and Snopes, often target different accounts: crowd fact-checkers tend to select posts from larger accounts with higher social influence, compared to expert fact-checkers [150].

Consensus in fact-checking: In traditional professional fact-checking, trained experts provide rigorous and authoritative assessments of claims, typically achieving high inter-rater agreement across fact-checkers from different organizations [152]. Conversely, community-based fact-checking relies on the collective judgment of lay contributors. The rationale is to identify misleading content by harnessing the wisdom of crowds [283]. Experimental studies support this premise: the aggregated ratings of even relatively small, politically balanced crowds can approach the accuracy of professional fact-checkers [46, 54].

However, the fact that crowds *can* produce accurate judgments does not guarantee that they *will* [52]. Several challenges may compromise the integrity of community-based fact-checking systems. These include intentional manipulation or gaming of the system [379], limited cognitive engagement when evaluating content [380], and motivated reasoning that shapes how users evaluate content [381]. These behaviors create opportunities for users to flag or downvote fact-checks not because they are inaccurate, but because they conflict with their ideological preferences or objectives. These challenges are further amplified by the high levels of political polarization on social media [382–384]. For example, contributors on community notes are more likely to rate fact-checks written by counter-partisans as unhelpful [50, 173]. To mitigate such polarization, X introduced a bridging-based algorithm that features community notes perceived as helpful across ideologically diverse subgroups [67]. Previous research has confirmed the effectiveness of the algorithm in surfacing high-quality community notes for display [67, 70]. In this context, high-quality sourcing plays a key role in fostering cross-partisan agreement and enabling the algorithm to identify notes with broad support [288, 385, 386].

Research gap. Different from professional fact-checking, where expert assessments are statically displayed on the corresponding misleading posts, the Community Notes system introduces a dynamic rating mechanism in which note status is continuously re-evaluated as new ratings arrive. Although the platform’s bridging-based algorithm has been shown to effectively surface helpful and broadly supported notes [67, 70], prior work has focused almost entirely on the *initial* selection of notes for display. Little is known about what happens *after* a note becomes visible. Because note status is dynamic, visibility itself may reshape subsequent evaluations (Fig. 8.1), potentially reinforcing agreement among like-minded users, provoking partisan backlash, or inviting strategic downvoting. Understanding these post-display rating dynamics is therefore crucial for improving algorithmic design and ensuring that community-based fact-checking systems remain both effective and resilient.

	(1) Displayed notes	(2) Stable	(3) Disappeared
#Notes	43,782	30,580 (69.8%)	13,202 (30.2%)
#Writers	19,260	14,257	8,338
Start date	12/11/2022	12/11/2022	12/11/2022
End date	06/27/2024	06/27/2024	06/16/2024
Posts with community notes			
#Posts	35,265	25,437 (72.1%)	11,901 (33.7%)
#Authors	13,553	10,627	5,758
Start date	12/11/2022	12/11/2022	12/11/2022
End date	06/10/2024	06/10/2024	06/10/2024
Ratings to community notes			
#Ratings	8,014,229	5,363,204 (66.9%)	2,651,025 (33.1%)
#Raters	541,833	508,994	390,285
Start date	12/11/2022	12/11/2022	12/11/2022
End date	07/14/2024	07/14/2024	07/13/2024

Table 8.1: Dataset overview. Columns (1)–(3) show statistics at note, post, and rating levels for displayed notes, stable notes, and disappeared notes, respectively.

8.3 Data and Methods

8.3.1 Dataset

To examine the stability of helpful community notes after their initial display, we use a large-scale dataset collected by a previous study that investigated the effectiveness of community fact-checking in reducing the spread of misleading posts [76]. The dataset comprises community notes and their associated source posts over the observation period from the public roll-out of the Community Notes program on November 11, 2022, to July 14, 2024. Specifically, it includes 437,396 community notes that had received 35,081,488 ratings from 583,285 raters, linked to 233,908 posts on $\mathbb{X}$.

Out of all notes in the dataset, only 43,782 (10.0%) were rated helpful after leaving the initial “Needs More Ratings” status and were displayed on their corresponding posts (see Suppl. S8.1 for details). However, 13,202 (30.2%) of displayed notes disappeared in later status updates. Thus, although only a small fraction of notes are ever displayed, a substantial portion of those that did appear were later removed. This pattern underscores the importance of examining the rating dynamics following the initial display of community notes. A summary of the dataset is presented in Table 8.1.

8.3.2 Reproduction of Note Selection Algorithm

The note selection algorithm is open-source and released by the Community Notes team on GitHub [387]. It is developed based on Matrix Factorization (MF) [67]. Specifically, all ratings are structured into a sparse note-rater matrix, each rating is predicted as:

$$\hat{r}_{un} = \mu + i_u + i_n + f_u \cdot f_n, \quad (8.1)$$

where $f_u \cdot f_n$ is the dot product of rater (user) vector f_u and note vector f_n . Additionally, μ is a global intercept term, accounting for an overall propensity of raters to rate notes “helpful” vs. “not helpful”; i_u is the rater’s intercept term, accounting for a user’s leniency in rating notes

“helpful”; and i_n is the note’s intercept, accounting for idiosyncratic “helpfulness” of the note beyond that explained by rater viewpoints and leniency, *i.e.*, the global helpfulness score assigned to the corresponding community note. Notably, the algorithm minimize a regularized least squared error loss function to estimate model parameters:

$$\min_{\{i,f,\mu\}} \sum_{r_{un}} (r_{un} - \hat{r}_{un})^2 + \lambda_i(i_u^2 + i_n^2 + \mu^2) + \lambda_f(\|f_u\|^2 + \|f_n\|^2), \quad (8.2)$$

where the regularization on the intercept terms, *i.e.*, $\lambda_i = 0.15$, is five times higher than the regularization on the factor terms, *i.e.*, $\lambda_f = 0.03$. Therefore, the algorithm is optimized toward explaining as much variation in the ratings as possible using rater and note factors, with the high estimated i_n indicating that the corresponding note is perceived helpful across diverse rater representations, *i.e.*, rater factor vectors. Additionally, to avoid overfitting, the note selection algorithm uses one-dimensional rater and note factors.

We use the version of the algorithm source code available at the time of data collection to calculate note intercepts and note factors for our analysis. The estimated *Core note intercept* reflects the perceived helpfulness of a community note, while the estimated *Core note factor* captures the polarization in its ratings [67] (see distribution and validation of these measures are provided in Suppl. S8.2).

8.3.3 Matrix Factorization for Rater Similarity

Given that the note selection algorithm emphasizes agreement among raters with diverse perspectives, it is essential to quantify the similarity between raters based on their past rating behaviors. Notably, the rater factor used by the Community Notes algorithm is intentionally one-dimensional to avoid overfitting on the initially small set of contributors, though higher-dimensional representations are expected as the contributor base grows [67]. In our study, given the large number of contributors and our focus on estimating nuanced rater representations rather than reducing overfitting, we adapt a matrix factorization algorithm that increases the dimensionality of both note and rater factors to 200. Additionally, to ensure sufficient information for model training, we restrict the dataset to notes that have received at least five ratings and raters who have rated at least five notes. The resulting dataset includes 34,596,349 ratings submitted by 406,181 raters on 372,980 notes.

Since our goal is to accurately estimate latent rater representations from the observed ratings without emphasizing generalization beyond the existing dataset, we omit the regularization terms and use a mean squared error loss function to optimize model fit:

$$\min_{\{i,f,\mu\}} \frac{1}{R} \sum_{r_{un}} (r_{un} - \hat{r}_{un})^2. \quad (8.3)$$

After 20 epochs of training, the loss converges to a stable level (see Fig. S8.2a, Suppl. S8.3), and the model demonstrates strong performance on the observed ratings, achieving a weighted $F1$ score of 0.962, a weighted precision of 0.969, and a weighted recall of 0.955.

Validation of the model estimates. We validate the reliability of the parameter estimates from the adapted MF model by comparing them with the corresponding parameters estimated by the note selection algorithm:

note factors and note intercepts. Given that the note factors in the adapted MF model are 200-dimensional vectors, we apply PCA to reduce their dimensionality for comparison with one-dimensional Core note factors in the note selection algorithm. The Core note factors and the PCA note factors have a significant positive linear relationship ($coef. = 2.790, p < 0.001$) with $R^2 = 0.792$ (see Fig. S8.2b, Suppl. S8.3). Additionally, we examine the distributions of estimated note intercepts, *i.e.*, indicator of helpfulness, in the note selection models and the adapted MF model. We find three clearly separated clusters indicating notes with the statuses of Currently Rated Helpful, Needs More Ratings, and Currently Rated Not Helpful (see Fig. S8.2c). This suggests that the adapted MF model produces note factors and intercepts that are highly consistent with the estimates from the note selection algorithm, effectively preserving the distinctions among different note statuses.

Similarity among raters. We calculate the similarity between raters that submitted ratings on the same note based on the 200-dimensional rater representations produced by the adapted MF model. Specifically, for each note, we examine all rater pairs and restrict our analysis to pairs that have co-rated at least five notes. We distinguish between rater pairs that assigned the same helpfulness level to a note and those that assigned opposite helpfulness levels. In total, this results in 78,166,385 rater pairs. We calculate the similarity for each rater pair and examine the distributions of similarities for rater pairs with same ratings on the same notes (see Fig. S8.2d; mean = 0.136, std. = 0.097) and rater pairs with the opposite ratings on the same notes (see Fig. S8.2e; mean = -0.097, std. = 0.094). Additionally, we observe a clear bimodal pattern in the overall distribution of rater similarities across all co-rating pairs (Fig. S8.2f), suggesting the presence of distinct groups of raters with different viewpoints. Therefore, in the subsequent analysis of the relationship between note writers and raters, we categorize raters as having *similar viewpoints* to a note writer if their similarity scores exceed ($>$) 0.136, and as having *opposite viewpoints* if their similarity scores are below ($<$) -0.097, based on their previous co-rating histories. Raters whose similarity scores to note writers lie between these thresholds are categorized as *general raters*.

8.3.4 Estimation Models

In this study, we employ a series of regression models and conduct empirical analyses based on the observational dataset to address our research questions. To address RQs 1–2, we specify a multilevel mixed-effects logistic regression model:

$$\text{logit}(P(Y_{ij})) = \beta_0 + \beta_1' X_i^{\text{Post}} + \beta_2' X_j^{\text{Poster}} + u_j + v_{i(j)}, \quad (8.4)$$

where Y_{ij} indicates whether the community note for post i from its author j disappears after its initial display ($= 1$) or not ($= 0$), and β_0 is the intercept. The vector X_i^{Post} represents variables related to post i , and the vector X_j^{Poster} represents variables related to post author j . Details on the post characteristics and author attributes included in the dataset are provided in Suppl. S8.4. Additionally, given that each post author can have multiple posts annotated by community notes, we incorporate multilevel random effects at both the post author level and post level. Specifically, u_j captures the random effect at the post author level, and $v_{i(j)}$ indicates the random effect at the post level, which is nested within post author level.

To estimate the changes in rating count and rating leaning after the initial display of community notes and to address RQs 3–4, we adopt an interrupted time series design specified as following:

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 D_i + \beta_3 D_i \times T_i + \gamma_i + \epsilon_i, \quad (8.5)$$

where Y_i represents either the *rating count* or the *rating leaning* for note i at time T_i . Here, a positive rating leaning (upvotes – downvotes) suggests that contributors prefer to rate notes as helpful, while a negative rating leaning suggest that contributors prefer to rate a note as not helpful. The variable T_i indicates the number of 15-min intervals (referred to as “quarters”) relative to the initial display of the community note (negative values indicate periods before display, and positive values indicate periods after display). The dummy variable D_i is a binary indicator equal to 1 for the period after the note’s initial display and 0 for the period before display. The interaction term $D_i \times T_i$ indicates the number of quarters since the initial display of the community note, while equal to 0 before display. Therefore, the estimate of β_2 captures the immediate effect at the display intervention, and the estimate of β_3 captures the sustained effect of the display intervention. Additionally, γ_i accounts for note-specific fixed effects, and ϵ_i is the error term.

Following previous research [223], we estimate a Poisson regression model with robust standard errors for the rating count to account for the non-negative and discrete nature of count data. For rating leaning, which is a continuous variable ranging from negative (more not-helpful ratings) to positive (more helpful ratings), we estimate an OLS linear regression model with robust standard errors. Notably, we also cluster robust standard errors at the note level to account for potential within-note correlations over time [77].

8.4 Results

8.4.1 Note Disappearance (RQ1 & RQ2)

We start by assessing the stability of the helpful status of community notes and the factors that are associated with their disappearance following initial display. As shown in Table 8.1, 30.2% (13,202) of displayed community notes later disappeared, indicating a substantial rate of removal. Figure 8.2 presents the estimated effects on the likelihood of note disappearance.

With respect to topic categories, community notes on posts about the economy (*coef.* = -0.220 , $p < 0.001$) or science (*coef.* = -0.376 , $p < 0.001$) are significantly less likely to disappear compared to notes on posts without these topics. Conversely, notes on posts related to health (*coef.* = 0.154 , $p < 0.01$) or politics (*coef.* = 0.169 , $p < 0.001$) are significantly more likely to disappear compared to notes on posts without corresponding topics. We find no significant relationship between note disappearance and either positive or negative sentiment in the source post; however, the presence of media elements is associated with a reduced likelihood of disappearance (*coef.* = -0.302 , $p < 0.001$). Author-level characteristics further shape disappearance patterns. Notes displayed on posts from high-influence users, *i.e.*, those with verified status (*coef.* = 0.202 , $p < 0.001$), older accounts (*coef.* = 0.209 , $p < 0.001$), or larger numbers of followers (*coef.* = 0.144 , $p < 0.01$), are more likely to disappear than notes on posts from low-influence users. In contrast, notes attached to posts from

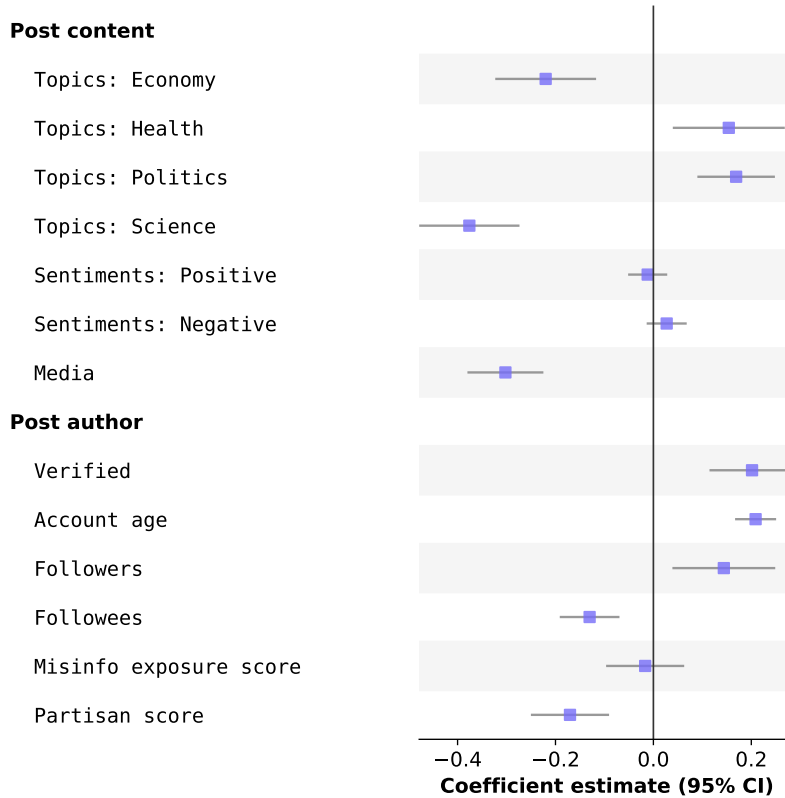


Figure 8.2: Estimation results for the likelihood of note disappearance. Shown are coefficient estimates with 95% Confidence Intervals (CIs). Random effects at both the author level and the post level are included during estimation.

authors with more followees ($coef. = -0.130, p < 0.001$) or right-leaning partisanship ($coef. = -0.170, p < 0.001$) are less likely to disappear than those from authors with fewer followees or left-leaning partisanship.

8.4.2 Changes in Rating Count and Leaning (RQ3)

Changes in rating count. We examine the rating counts over an observation window spanning -16 to 16 quarters relative to the initial display of community notes and estimated the effect of note display using an ITS design. As shown in Fig. 8.3a, note display produces a sharp immediate increase of 39.3% (95% CI: [35.7%, 43.0%]; $p < 0.001$) in the number of ratings compared to the pre-display period. Notably, this immediate effect is substantially larger for disappeared notes (89.5%, $p < 0.001$; 95% CI: [80.3%, 99.2%]) than for stable notes (17.3%, $p < 0.001$; 95% CI: [14.3%, 20.4%]). Additionally, a relatively modest yet statistically significant sustained increase of 3.6% per quarter (95% CI: [3.3%, 3.9%]; $p < 0.001$) is observed over time, which is similar across disappeared notes (3.4%, $p < 0.001$; 95% CI: [2.8%, 4.0%]) and stable notes (3.8%, $p < 0.001$; 95% CI: [3.5%, 4.2%]). Taken together, these findings indicate that the primary difference in the effect of note display between disappeared notes and stable notes lies in the magnitude of the immediate post-display response. Accordingly, in the subsequent analyses, we focus on the immediate effect following note display, which captures the direct impact of the display intervention.

Changes in rating leaning. Fig. 8.3b shows that the rating leaning increases by 0.255 (95% CI: [0.225, 0.286]; $p < 0.001$) following note display. This suggests that the increased ratings observed in Fig. 8.3a are primarily driven by upvotes exceeding downvotes. Particularly, the

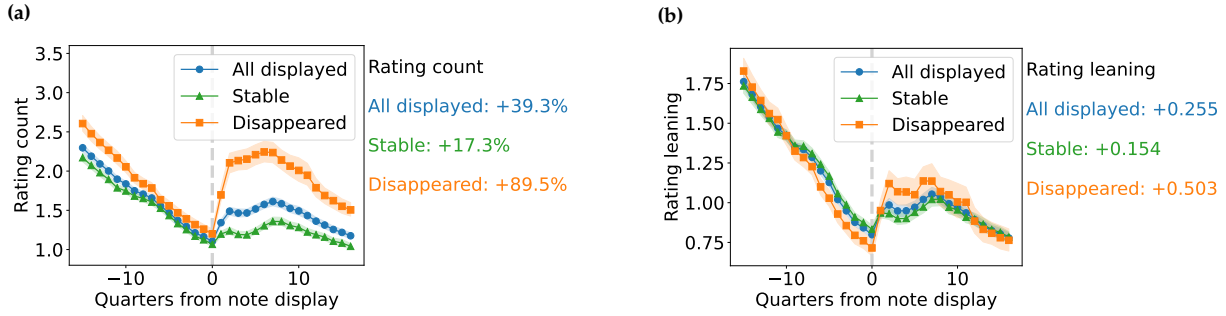


Figure 8.3: Changes in rating count and rating leaning after the initial display of community notes. (a) Average rating count per quarter (15-minute interval) from -16 to 16 quarters relative to note display. (b) Average rating leaning per quarter over the same window. The results are shown for all displayed notes, stable notes, and disappeared notes. Error bands represent 95% CIs.

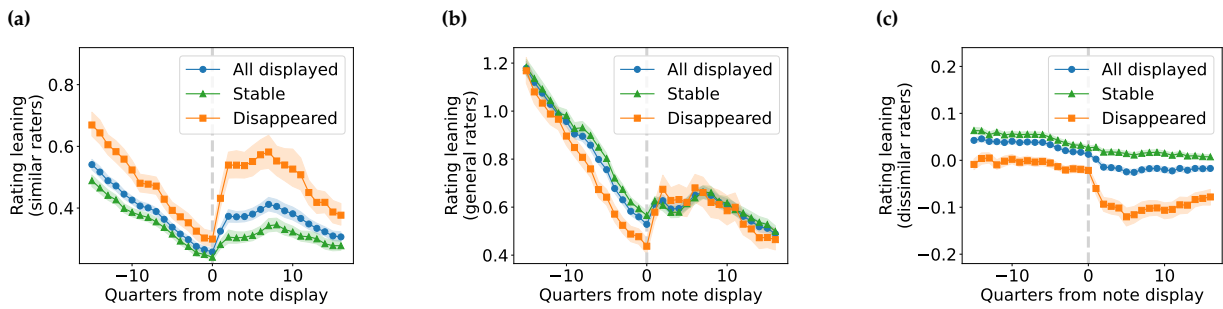


Figure 8.4: Discrepancies in rating leaning between similar and dissimilar raters. (a) Average rating leaning per quarter (15-minute interval) for similar raters from -16 to 16 quarters relative to note display. (b) Average rating leaning per quarter for general raters over the same window. (c) Average rating leaning per quarter for dissimilar raters over the same window. The results are shown for all displayed notes, stable notes, and disappeared notes. Error bands represent 95% CIs.

increase in rating leaning for disappeared notes (0.503 , $p < 0.001$; 95% CI: $[0.428, 0.577]$) is significantly larger than that for stable notes (0.154 , $p < 0.001$; 95% CI: $[0.124, 0.185]$).

Sensitivity analysis. We further examine changes in rating leaning across post content categories and author characteristics (see Suppl. S8.5). We find that the estimated effect sizes vary across subgroups, but almost all are significantly positive (with the exception of health-related posts). This suggests that note display consistently attracts more upvotes than downvotes across a wide range of post content types and author attributes.

8.4.3 Similar vs. Dissimilar Raters (RQ4)

Although community notes generally receive more upvotes than downvotes following their initial display, many still disappear. As shown earlier, disappeared notes tend to attract a greater number of ratings and exhibit higher rating leaning after display compared to stable notes. Given that the note selection algorithm emphasizes agreement among raters with diverse viewpoints, it is possible that the increased ratings following note display are biased or polarized, thereby increasing the likelihood of disappearance. To investigate this, we examine differences in rating leaning between raters whose past rating behavior is similar to that of the note writer and those who are dissimilar.

In our dataset, a total of 1,402,344 ratings fall within the observation window spanning -16 to 16 quarters from note display, of which 371,277

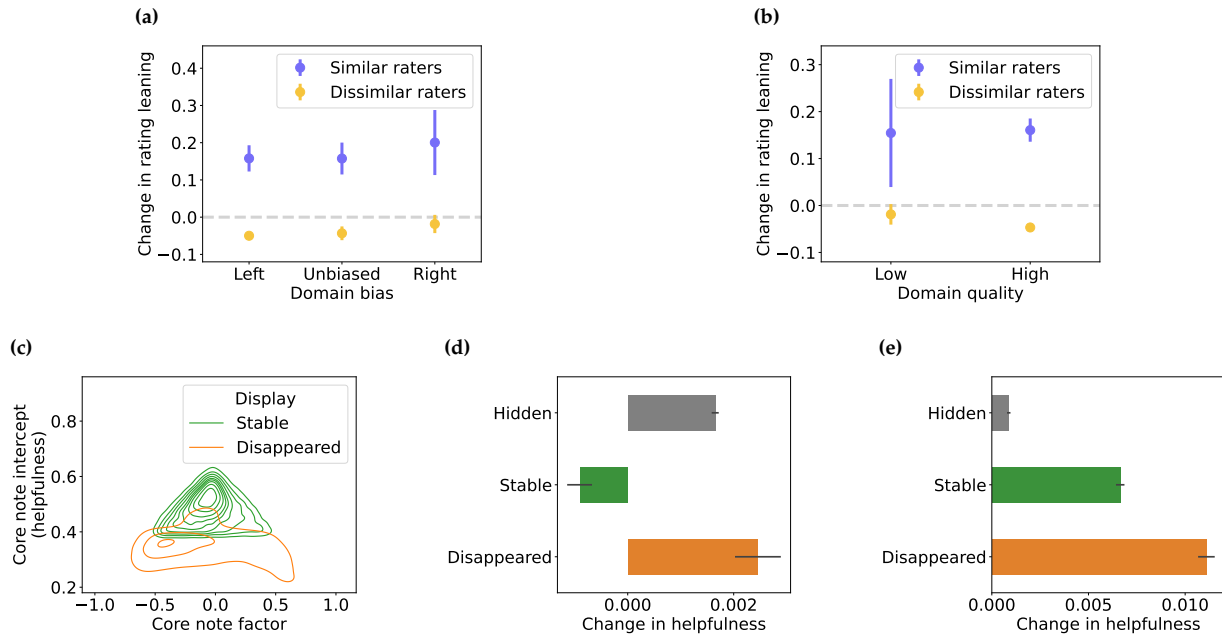


Figure 8.5: Evaluation of polarized ratings and counterfactual analysis of the note selection algorithm. (a) Estimated changes in rating leaning for similar and dissimilar raters across levels of political bias in external domains cited in community notes. (b) Estimated changes in rating leaning for similar and dissimilar raters across high- and low-quality external domains. (c) Distributions of estimated Core note intercepts (helpfulness) and estimated Core note factors (polarization) from the note selection algorithm, shown for disappeared notes and stable notes. (d) Changes in helpfulness when ratings from similar and dissimilar raters are excluded. (e) Changes in helpfulness when ratings from dissimilar raters are excluded. Results are shown for disappeared notes, stable notes, and notes that have never been displayed. The error bars represent 95% CIs.

(26.5%) originate from similar raters, and 110,103 (7.9%) from dissimilar raters. Thus, viewpoint-aligned and viewpoint-opposed raters constitute a minority, but potentially an influential subgroup relative to the majority of general raters (65.6%).

Fig. 8.4a shows that similar raters exhibit a sharp increase in rating leaning following note display ($coef. = 0.148, p < 0.001$; 95% CI: [0.133, 0.163]). The increase in rating leaning for disappeared notes ($coef. = 0.298, p < 0.001$; 95% CI: [0.262, 0.334]) is significantly larger than that for stable notes ($coef. = 0.087, p < 0.001$; 95% CI: [0.073, 0.101]). General raters exhibit a similar pattern (Fig. 8.4b), showing a significant increase in rating leaning following note display ($coef. = 0.139, p < 0.001$; 95% CI: [0.119, 0.158]). Likewise, the increase for disappeared notes ($coef. = 0.287, p < 0.001$; 95% CI: [0.240, 0.334]) is significantly stronger than that for stable notes ($coef. = 0.070, p < 0.001$; 95% CI: [0.059, 0.098]). In contrast, Fig. 8.4c shows that the rating leaning for raters who have dissimilar viewpoints with note authors drops significantly following note display ($coef. = -0.032, p < 0.001$; 95% CI: [-0.037, -0.027]). This decrease is more pronounced for disappeared notes ($coef. = -0.083, p < 0.001$; 95% CI: [-0.096, -0.069]) compared to stable notes ($coef. = -0.011, p < 0.001$; 95% CI: [-0.015, -0.008]).

Taken together, these results reveal a pattern of polarized rating behavior following note display: supportive ratings increase among similar raters while agreement decreases among dissimilar raters, with polarization substantially stronger for notes that ultimately disappear. This pattern is robust across post content categories and author characteristics, with especially pronounced polarization for political posts and left-leaning authors (see Fig. S8.3b, Suppl. S8.5).

8.4.4 Counterfactual Analysis

The observed rating polarization between similar and dissimilar raters may have two competing implications: (i) it may prevent the continued display of potentially problematic or low-quality notes through cross-viewpoint disagreement, or (ii) it may reflect coordinated or ideologically motivated rating behaviors that undermine consensus-based evaluation. Given that the overall rating trend for displayed community notes is strongly positive (Fig. 8.4b), the latter explanation, *i.e.*, ideologically driven rating behaviors, appears more plausible. To further validate this, we assess the political bias and information quality of external domains cited in community notes (see Suppl. S8.6) and conduct counterfactual analyses by re-running the note selection algorithm after excluding potentially polarized ratings (*i.e.*, those from similar and dissimilar raters).

As shown in Fig. 8.5a, similar raters exhibit consistent positive rating leanings for community notes with left-leaning ($coef. = 0.158, p < 0.001$; 95% CI: [0.123, 0.193]), unbiased ($coef. = 0.157, p < 0.001$; 95% CI: [0.115, 0.200]), and right-leaning domains ($coef. = 0.200, p < 0.001$; 95% CI: [0.113, 0.288]). In contrast, dissimilar raters show negative rating leanings for notes with left-leaning ($coef. = -0.050, p < 0.001$; 95% CI: [-0.060, -0.040]) and unbiased domains ($coef. = -0.043, p < 0.001$; 95% CI: [-0.062, -0.025]), while their evaluations of notes citing right-leaning domains are not statistically distinguishable from zero. A similar pattern emerges when examining domain quality (Fig. 8.5b): similar raters show positive rating leaning for notes citing both low-quality ($coef. = 0.155, p < 0.001$; 95% CI: [0.039, 0.270]) and high-quality domains ($coef. = 0.161, p < 0.001$; 95% CI: [0.136, 0.185]); conversely, dissimilar raters show negative rating leaning, especially for notes with high domain quality ($coef. = -0.047, p < 0.001$; 95% CI: [-0.056, -0.038]). Together, these findings suggest the presence of polarized rating behaviors, *e.g.*, towards notes citing left-leaning or high-quality domains.

Fig. 8.5c further shows a clear separation between disappeared and stable notes: disappeared notes are associated with lower estimated helpfulness and higher inferred polarization (farther from zero). To assess whether the polarization in rating behavior between similar and dissimilar raters contributes to the disappearance of community notes, we rerun the note selection algorithm based on ratings from only general raters (*i.e.*, excluding ratings originated from similar and dissimilar raters after note display). Compared with the original helpfulness scores (based on all ratings), counterfactual helpfulness scores increase substantially for disappeared notes with a mild decrease for stable notes (Fig. 8.5d). This pattern suggests that dissimilar raters reduce consensus for notes that otherwise would appear more helpful.

Because rating patterns of similar raters are largely consistent with those of general raters, while dissimilar raters show divergent and often oppositional evaluations, we further isolate the contribution of dissimilar raters by excluding only their ratings. As shown in Fig. 8.5e, removing dissimilar raters leads to the largest increases in estimated helpfulness for disappeared notes, followed by stable notes. Overall, the counterfactual analysis supports the interpretation that the increased polarization in ratings following note display, especially negative evaluations from a small group of dissimilar raters, significantly reduce cross-group consensus and contribute directly to note disappearance.

8.5 Discussion

Here, we examine the stability of community notes displayed on X and the vulnerability of its consensus-based note selection algorithm. We find that a substantial share of notes (30.2%) that initially meet the platform's display criteria (*i.e.*, were rated as helpful) later disappear. More critically, disappearance is strongly associated with posts on high-stakes topics (*e.g.*, health, politics) and with posts from high-influence authors (*e.g.*, verified accounts with many followers). We also observe an asymmetry by political orientation: notes displayed on posts from left-leaning authors are more likely to disappear compared to those from right-leaning authors.

Disappearance mechanism of community notes. Two mechanisms may explain the disappearance of community notes: (i) the notes are genuinely unhelpful or inaccurate, or (ii) the display of community notes may amplify ideologically biased or strategically coordinated rating behaviors that push notes below the display threshold after they appear. Prior work demonstrates that the note selection algorithm is generally effective at surfacing high-quality notes, making it unlikely that widespread disappearance is driven primarily by accuracy issues. Our analysis of rating patterns supports this interpretation. After appearing on a post, notes typically receive more upvotes than downvotes—a pattern that persists even after excluding similar and dissimilar raters who may possess systematic biases. However, we identify systematic polarization in rating behavior between similar and dissimilar raters following note display. Similar raters become more supportive, dissimilar raters become less supportive, and this divergence is particularly pronounced for notes that ultimately disappear. Our counterfactual analysis indicates that this post-display polarization in ratings, particularly from dissimilar raters, plays a substantial role in driving note disappearance. Together, our findings indicate that note disappearance is less a reflection of note quality and more a consequence of ideologically polarized or strategically motivated rating behavior.

Vulnerability of consensus-based algorithms. While the consensus-based note selection algorithm on X is designed to surface notes that are perceived as helpful by a broad and ideologically diverse set of raters, emerging evidence suggests that consensus is difficult to achieve: the proportion of notes deemed “helpful” and subsequently displayed relative to all community notes over time remains low (around 10%) and is even declining [76, 377, 388, 389]. Although politically balanced crowds have demonstrated the capacity to identify misinformation effectively at scale [46, 54, 281, 390], community-based fact-checking systems remain fragile in polarized social media environments, where partisan disagreement, motivated reasoning, and coordinated rating behaviors can destabilize consensus-based moderation [50, 68, 376, 377, 381, 391, 392]. For instance, users are more likely to provide negative ratings to evaluations from counter-partisans [50, 145]. Therefore, many factually accurate but politically sensitive notes often fail to meet the cross-group consensus threshold and remain unpublished, reflecting systematic suppression of information that is credible but polarizing [377].

Our findings extend this literature by showing that the *display* of community notes itself can intensify polarization in subsequent ratings. Even when only a small fraction of ratings exhibit systematically oppositional or strategically coordinated behavior, their concentrated negative ratings disproportionately contribute to note suppression [376]. This dynamic

not only undermines the persistence of currently displayed notes but may also reinforce rating biases in subsequent update rounds, further reducing the proportion of notes that successfully reach publication. More broadly, such post-display polarization may undermine the long-term trustworthiness and effectiveness of community-based fact-checking. While displayed notes can increase user trust and engagement with misinformation [56, 76, 298, 393], polarized ratings can cause high-quality notes to disappear, potentially eroding their effectiveness. Furthermore, the scalability and timeliness of Community Notes system rely on efficient aggregation of crowd ratings [47, 389, 394], yet biased or coordinated post-display ratings threaten this efficiency. Our findings therefore highlight the importance of mitigating post-display polarization in rating behavior to ensure the efficacy and sustainability of community-based fact-checking systems.

Practical implications. Our study underscores the importance of carefully designing post-display update mechanisms and interventions, as display itself can amplify polarization in subsequent evaluations. This has several practical implications for platforms relying on consensus-based moderation and community-based fact-checking. First, platforms may incorporate safeguards to detect and exclude potentially coordinated or manipulated rating behaviors, thereby improving the resilience of consensus-based algorithms against bias and manipulation [395]. Second, platforms could consider adopting a multi-layer fact-checking ecosystem that incorporates contributions from both crowd and expert fact-checkers, with experts taking the lead on high-stakes or politically charged content [45, 48, 286, 390]. Such hybrid arrangements may enhance accuracy, reduce vulnerability to polarized rating behaviors, and ensure more stable fact-check visibility on social media platforms. Finally, advancements in Large Language Models (LLMs) offer promising opportunities to support and strengthen community-based fact-checking [45, 396, 397]. For instance, LLMs can assist contributors with evidence retrieval, note generation, and information synthesis [213, 396–399], and may also help reduce bias in rating processes by providing structured evaluation assistance [400].

Limitations and future work. Our results point to several promising directions for future research. First, while we document substantial post-display polarization in rating dynamics, the current data do not permit a clear separation between individual bias, partisan disagreement, and coordinated manipulation. Future studies combining behavioral traces with survey or experimental data could clarify the mechanisms underlying polarized evaluations of community fact-checks. Second, the generalizability of our findings is conditioned by the specific design features of Community Notes on X, especially the platform’s evolving criteria for note display. As other platforms experiment with community-based fact-checking, comparative analyses could identify which design elements enhance the stability of evaluations and protect against coordinated manipulation. Finally, our study relies on a snapshot of the Community Notes dataset, which may omit notes or ratings that were later removed or edited. Longer-term archival efforts that provide more complete historical traces would enable richer analyses of the lifecycle of notes and how contributor behavior evolves over time.

Abstract: X’s Community Notes is a crowdsourced fact-checking system. To improve its scalability, X introduced the “Request Community Note” feature, enabling users to solicit fact-checks from contributors on specific posts. Yet, its implications for the system—what gets checked, by whom, and with what quality—remain unclear. Using 98,685 requested posts and their associated notes, we evaluate how requests shape the Community Notes system. We find that requested posts with higher GPT-estimated misleadingness and from authors with greater misinformation exposure are more likely to receive notes. Conversely, requested political posts receive fewer notes than non-political ones. We also observe partisan asymmetries: posts from Republicans are more likely to receive notes than those from Democrats. Although only 12% of requested posts receive request-fostered notes from top contributors, these notes are rated as more helpful and less polarized than others, partly reflecting top contributors’ selective fact-checking of misleading posts. Our findings highlight both the limitations and promise of requests for scaling high-quality community-based fact-checking.

Publication
Y. Chuai, S. Zhang, Z. Wang, X. Yi, M. Mosleh, and G. Lenzini, ‘Request a note: How the request function shapes X’s Community Notes system,’ in <i>Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems</i> , 2026, pp. 1–22.

9.1 Introduction

Designing effective countermeasures to identify falsehoods and curb the spread of misleading posts (*i.e.*, posts carrying misinformation) on social media platforms is still a challenge. Years ago, the social media platform X (formerly Twitter) adopted *Community Notes*, a crowdsourced fact-checking system designed to address misleading posts and provide users with helpful additional context.¹ This initiative marks the first large-scale implementation of community-based fact-checking by a major social media platform [45, 74]. A subset of users—called *contributors*—are periodically admitted into the program if they meet specific eligibility requirements, such as having no recent violations of X’s rules. Contributors begin by rating the helpfulness of existing notes, and only after demonstrating reliability can they unlock the ability to write notes themselves. All notes are evaluated by a note selection algorithm, and then those deemed helpful by the algorithm are displayed directly on the relevant posts [66]. Inspired by this program, other social media providers, including YouTube, Meta, and TikTok, are beginning to explore similar community-based moderation models [367, 403, 404]. For example, Meta is piloting the Community Notes program across Facebook, Instagram, and Threads in the U.S., while TikTok plans to offer a similar feature called Footnotes to provide context and corrections for misleading posts. Given the rapid expansion of community-based fact-checking system, it is crucial to understand how the system functions in the real world.

9.1 Introduction 167

9.2 Background and Related Work 169

9.3 Data and Methods . . . 170

9.3.1 Data Collection 170

9.3.2 Characteristics of Post Content 171

9.3.3 Characteristics of Post Authors 173

9.3.4 External Domain References in Community Notes 174

9.3.5 Identification of Top Writers 174

9.3.6 Replication of Note Selection Algorithm . . 175

9.3.7 Estimation Models . . . 175

9.4 Empirical Results 176

9.4.1 Likelihood of Receiving Community Notes . . . 176

9.4.2 Request Timing and Request-Fostered Notes 178

9.4.3 Algorithmic Evaluation of Community Notes . . 180

9.4.4 Summary of Main Findings 185

9.5 Discussion 185

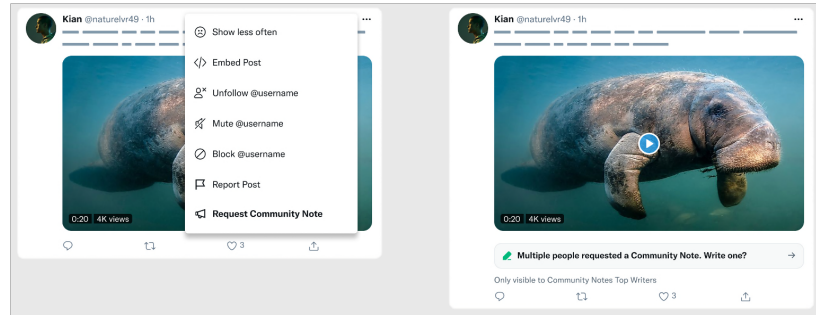
9.5.1 Research Implications . 185

9.5.2 Practical Implications . 187

9.5.3 Limitations and Future Work 188

1: Here, misinformation refers to false or misleading information that is contradicted by empirical evidence or inconsistent with common or expert consensus, without imposing assumptions on the intent or the format of the content [74]. It serves as an umbrella term including not only factually incorrect content but also disinformation and fake news—both deliberately fabricated to mislead others and/or cause harms [1, 74, 401]. Prior work further shows that factually accurate yet misleading content can be more influential than outright falsehoods [402]. In the Community Notes system, the criteria used for labeling misleading posts align closely with our definition of misinformation (see details in Suppl. S9.1).

Figure 9.1: The introduction of “Request Community Note” feature on X, sourced from the official Community Notes website [405]. It allows users to request a note on a post they believe would benefit from additional context (see the left part of the figure). When a sufficient number of requests are submitted, a subset of Community Notes contributors (i.e., top writers) will see an alert below the corresponding requested post, and can choose to propose a note (see the right part of the figure).



Recent work has shown that X’s Community Notes program can reliably attach helpful notes to the corresponding misleading posts [67, 70]. Moreover, community fact-checks are perceived as more trustworthy than expert fact-checks, can effectively reduce the spread of misleading posts, and even prompt authors to delete their problematic content [56, 67, 76, 77]. However, the Community Notes system, in its current implementation on X, still faces limitations, such as the small number of displayed notes relative to all generated notes and restrictions on who can author them [76].

To foster note writing and create a more interactive and responsible community, X recently implemented a key innovation in Community Notes system—the *Request Community Note* feature (as illustrated in Fig. 9.1), allowing broader registered users to signal posts they believe would benefit from additional context. This mechanism enables non-contributor users to participate in the moderation process while guiding contributors toward posts where notes are likely to be needed. This request feature is expected to increase the involvement and the interest of users in checking facts, which is crucial in the current media landscape, where misinformation can spread rapidly. However, it remains unclear whether this mechanism effectively improves the Community Notes system by helping more misleading posts receive notes and enhancing the quality of the resulting notes.

Research questions. In this study, we examine the newly introduced Request Community Note feature on X, tracing the process from the submission of requests to the evaluation of the associated notes. The workflow involves requestors flagging posts (*i.e.*, *requested posts*) for note writers, note writers selecting requested posts to propose community notes, and the note selection algorithm evaluating the generated notes. Notably, note writers retain autonomy: they can independently choose any post to fact-check and propose community notes, regardless of whether they see request alerts or not. Moreover, the note selection algorithm evaluates community notes along two key dimensions: *helpfulness*, indicating the consensus of perceived helpfulness among heterogeneous contributors, and *polarization*, capturing the degree of variation in ratings. Accordingly, we organize our investigation around three research questions (RQs) to explore how note writers select requested posts to fact-check, the extent to which their selections are influenced by requests, and how community notes fostered by user requests (*i.e.*, *request-fostered notes*) differ from other notes in helpfulness and polarization according to the note selection algorithm:

- **RQ1:** Among requested posts, which ones actually receive community notes, and how do their content and author characteristics shape this likelihood?

- **RQ2:** To what extent do user requests foster the creation of community notes, and how do they influence contributors' selection of posts?
- **RQ3:** Are community notes fostered by user requests more or less helpful and polarized than other notes, and what factors explain these differences?

Data and methods. To address our research questions, we collect all available Community Notes data and associated posts that had received at least five requests as of May 20, 2025. We then specify a logistic regression model to estimate the likelihood of a post receiving a community note based on a wide range of characteristics extracted from posts, such as topics and estimated misleadingness, and their authors, such as verified status and misinformation exposure scores [17] (RQ1). In addition, we analyze the extent to which requests foster the creation of community notes and how requests can influence the choices of note writers (RQ2). Finally, we replicate the note selection algorithm to evaluate request-fostered notes compared to other notes and explore factors affecting the evaluation results (RQ3).

Contributions. Overall, our study results in three main findings. (i) The likelihood that a requested post receives a community note is associated with the post's content and its author. For instance, posts related to entertainment or finance, containing claim-oriented content, with higher estimated misleadingness, or from accounts with higher misinformation exposure scores, are more likely to receive community notes. Our findings suggest that Community Notes contributors and requestors may have different selection patterns. (ii) With more than half (53.6%) of requested posts receiving community notes, only an estimated 12.1% of those are likely fostered by user requests, showing that the effect of requesting community notes in fostering note generation is, so far, limited. Nonetheless, requests tend to shift contributors' attention toward political content, while prioritizing posts with higher misleadingness (as measured by GPT-4.1 from OpenAI). (iii) Request-fostered notes generated by top writers are evaluated as more helpful and less polarized than other notes, a pattern that likely reflects writers' selective prioritization of requested posts with higher levels of misleadingness. To our knowledge, our study provides the first empirical evaluation of this novel request mechanism, offering insights into both the opportunities and challenges of scaling crowdsourced fact-checking systems.

9.2 Background and Related Work

The spread of misleading posts on social media platforms—including X (formerly Twitter), TikTok, and Facebook—remains a significant societal problem, with far-reaching and often harmful consequences [5, 33]. For example, the widespread misinformation during political elections is particularly alarming, as exposure to and belief in false narratives during elections can distort public opinion, erode trust in democratic institutions, and destabilize societies [4, 14, 33, 253]. Beyond electoral contexts, misinformation can exacerbate social divisions and intensify polarization in response to global pandemics and claim change [1, 33]. Given these multifaceted risks, social media providers face increasing pressure to adopt transparent content moderation policies and implement effective and scalable interventions to identify misleading posts, correct falsehoods, and ensure users have access to reliable information [5, 141, 263, 303].

Research has shown that Community Notes system can successfully identify misleading content, reduce its spread, and even pressure authors to delete problematic posts [67, 70, 76, 406]. Additionally, community-based annotations are often perceived as more trustworthy than professional fact-checks [56]. These encouraging findings suggest that Community Notes is promising in addressing misinformation on social media platforms and have sparked growing scholarly interest in understanding the dynamics of the system and identifying opportunities to refine and enhance its design [45]. For instance, despite the relatively large volume of notes generated by contributors, only a small fraction—around 10%—achieve sufficient support from diverse users and are displayed to users [76, 398]. At the same time, the display process is often not fast enough to intervene before misleading posts have already begun to spread widely [74, 76]. To address these limitations, recent work has explored two complementary directions.

First, researchers and practitioners have turned to technological augmentation, using Large Language Models (LLMs) to help scale human judgment by assisting contributors in drafting and refining notes [396]. For example, the Supernotes approach generates AI-synthesized summaries that integrate insights from multiple existing contributor notes with the goal of fostering consensus among diverse raters [398]. Second, X has expanded user participation by introducing the request function, which allows any user with a verified account to flag posts that they believe should receive a community note. This structural expansion extends the influence of the wider user base beyond the core contributor group, potentially increasing the coverage and responsiveness of the system. However, despite its significance, this new function has not yet been systematically studied, leaving open questions about its effects on content selection for fact-checking, note generation, and the quality of resulting notes. Addressing these questions represents the primary goal of this study.

9.3 Data and Methods

In this section, we first describe our data collection process, including the posts, requests, and associated community notes. We then outline the post content and author characteristics extracted for analysis, including GPT-based evaluations of post misleadingness as a proxy for misinformation risk. Next, we examine external source domains cited in community notes, assessing them in terms of information quality and political bias. Since request alerts are only visible to top writers, we also identify top note writers following the criteria provided on the official Community Notes website. Finally, we replicate the note selection algorithm to evaluate the helpfulness and polarization of notes, enabling us to compare request-fostered notes with those generated independently of requests.

9.3.1 Data Collection

To address our research questions, we downloaded the complete set of available Community Notes data—including requests, notes, ratings, and note status histories—on May 20, 2025, when X began releasing request data to the public [299]. Specifically, we collected 5,888,351 requests submitted by 1,689,152 unique users since the launch of the request

Table 9.1: Overview of requested posts. The three columns are for all requested posts, posts with community notes, and posts without community notes, respectively. The mean values for continuous variables (standard deviations in parentheses), percentages for dummy variables, or count numbers for count variables are reported.

	(1) All	(2) With notes	(3) Without notes
# Requested posts	98,685	52,862 (53.6%)	45,823 (46.4%)
# Post authors	22,915	15,575	12,779
<u>Post content characteristics</u>			
Sentiments: Positive	0.131 (0.244)	0.134 (0.246)	0.127 (0.241)
Sentiments: Negative	0.460 (0.335)	0.448 (0.335)	0.474 (0.335)
Topics: Politics	37.0%	32.4%	42.4%
Topics: Science & Technology	13.5%	12.8%	14.2%
Topics: Entertainment	26.9%	28.0%	25.6%
Topics: Finance & Business	32.6%	33.9%	31.2%
Content types: Claim	0.687 (0.318)	0.687 (0.324)	0.687 (0.312)
Content types: Opinion	0.489 (0.346)	0.475 (0.347)	0.506 (0.344)
GPT misleadingness	0.398 (0.318)	0.405 (0.324)	0.389 (0.310)
Media	65.4%	68.4%	62.0%
<u>Post author characteristics</u>			
Account types: Blue	72.5%	72.3%	72.6%
Account types: Business	6.6%	6.4%	6.7%
Account types: Government	7.2%	6.5%	7.9%
Account age	3,156.145 (1,997.934)	3,078.018 (1,998.710)	3,246.273 (1,993.259)
Followers	6,828,060	6,561,490	7,135,578
Followees	13,347	11,081	15,960
Misinformation exposure score	0.546 (0.139)	0.555 (0.143)	0.536 (0.135)
Partisan score	−0.035 (0.767)	0.023 (0.768)	−0.095 (0.761)

feature on July 18, 2024. In addition, the dataset includes 1,787,609 notes that have received 149,646,500 ratings since the introduction of Community Notes program.

Subsequently, we used the full-archive search endpoint of the X Pro API to collect the requested posts. In total, the submitted requests target 2,427,451 unique posts. However, according to the platform guidelines at the time of data collection, a post must receive at least 5 requests to be surfaced to top writers, indicating a threshold for which posts are eligible for community evaluation [405]. In our dataset, 154,090 posts meet this criterion, representing only 6.3% of all requested posts. In contrast, the requests submitted to these posts account for more than half (52.6%) of the total requests. This suggests that, while users submitted a large volume of requests, attention was concentrated on a small set of posts. Based on the corresponding IDs of these eligible posts, we successfully retrieved 133,351 posts, of which 98,685 (74%) are in English and were authored by 22,915 unique accounts. We restrict our analysis to the collected posts in English and focus on the request feature in the U.S. context. Next, we introduce how to extract the characteristics of posts and their authors for our subsequent analysis (see summary statistics in Table 9.1).

9.3.2 Characteristics of Post Content

Sentiments. Sentiments are an important factor that influences the dissemination of online content [10, 98]. We extract sentiments in the posts content using a state-of-the-art Twitter-roBERTa-base sentiment

2: HuggingFace

classification model (2022 updated), which is available on HuggingFace.² This model was trained on 124 million posts created from January 2018 to December 2021, and fine-tuned for sentiment analysis. It achieves superior predictive performance on the TweetEval benchmark, with a reported macro-recall of 0.737 on the sentiment classification task [350]. The model outputs probability estimates for three sentiment categories: positive, neutral, and negative, with the sum of the probabilities across these categories equaling to 1. On average, requested posts exhibit strongest negative sentiment (mean of 0.460), followed by neutral sentiment (mean of 0.409; $t(\text{negative}, \text{neutral}) = 28.651, p < 0.001$) and positive sentiment (mean of 0.131; $t(\text{negative}, \text{positive}) = 199.208, p < 0.001$). In this study, we focus on affectively charged content; thus, to avoid multicollinearity in our empirical regression analyses, we retain the probabilities of positive and negative sentiments for each collected post and omit the neutral probability.

Topics. Online misinformation often leverages sentiments to enhance its virality and tends to concentrate on specific topics—most notably, *e.g.*, politics [10, 14]. X has provided domain labels for each post in the collected dataset. Following prior work [74], we extract these domains from the requested posts and group them into four broad topics: Politics, Science & Technology, Entertainment, and Finance & Business. Notably, a single post may be associated with multiple topics. Politics (37%) is the most frequent topic in the requested posts, followed by Finance & Business (32.6%), Entertainment (26.9%), and Science & Technology (13.5%).

Content types and misleadingness. Traditional expert-based fact-checking and emerging LLM-based misinformation detection methods primarily aim to identify checkworthy, objective, and verifiable claims, subsequently evaluating their truthfulness [62, 73]. These approaches focus on factual claims and often overlook opinion-based content that may contain misleading interpretations [407]. In contrast, the Community Notes system relies on crowd consensus to assess the misleadingness of selected posts and surface community fact-checks, with the potential to address both factual inaccuracies and misleading opinions (see the misleading criteria in Suppl. S9.1). Given this, we distinguish between *factual claims*, which explicitly presents some verifiable facts, and *personal opinions*, which are more subjective content to express users' personal judgments (see full definitions in Suppl. S9.2). This distinction is important for understanding how requests and community notes select verifiable content and opinionated posts. Notably, recent critiques also highlight that a substantial portion of community notes have been attached to opinion-based posts and published for debate rather than moderation [389]. To assess whether requestors and contributors target content that is plausibly misleading, rather than merely content they disagree with, we analyze the misleadingness of requested posts.

Given that recent work demonstrates the comparable performance of LLMs, such as GPT, to human evaluations in text quality and their potential applications in large-scale empirical analysis [61, 63, 408], we use GPT-4.1, a state-of-the-art LLM model at the time of this study, to classify content types (claims vs. opinions) and to estimate the misleadingness of requested posts. To ensure that the model's outputs align with the scope of our analysis, we provide GPT a task-specific prompt incorporating the definitions and criteria for misleading information, factual claims, and opinions (see full prompt in Suppl. S9.2). The GPT-estimated scores for claim, opinion, and misleadingness for each post range from 0 to 1, with higher values indicating a stronger presence of the corresponding

attribute. Additionally, a post can contain both claims and opinions, and their estimated scores are not required to sum to 1.

To assess the reliability of GPT-generated annotations, we first randomly select 200 posts from the requested dataset and conduct a manual validation using a small crowd. Specifically, seventeen university-level student assistants are recruited and trained on the relevant definitions and criteria to annotate whether each assigned posts contain claims, opinions, and misleading information. Each post is independently reviewed by a minimum of six assistants to reduce individual-level noise and ensure robust aggregation, with final annotations determined by averaging their ratings. For comparison, we transform the GPT-estimated scores into binary classes using a threshold of 0.5, and compare the GPT-generated labels with the assistant-annotated labels. We find that the GPT-based approach demonstrates strong performances for claims (*Accuracy* = 0.840, *Weighted F1* = 0.838), opinions (*Accuracy* = 0.805, *Weighted F1* = 0.810), and misleadingness (*Accuracy* = 0.800, *Weighted F1* = 0.800). These results suggest that the GPT-generated outputs are comparable to the aggregated human judgments, supporting their use in our large-scale analysis (see its validation with expert fact-checks from PolitiFact in Suppl. S9.3). Overall, the requested posts are more likely to contain claims (mean of 0.687), compared to opinions (mean of 0.489; $t = 111.269$, $p < 0.001$). Additionally, the GPT-estimated misleadingness is, on average, 0.398 in the requested posts.³

Media posts. Media elements, such as images or videos, can increase user engagement with associated posts, making media-based misinformation potentially more viral and harmful than text-only content [74, 409]. Additionally, the Community Notes team on X implemented a media note feature, which allows helpful media notes to be displayed across all posts that contain the same media [410]. This feature can help curb the spread of misleading posts at the very beginning if they contain the same misleading media annotated by existing community notes, thereby increasing the responsiveness of the Community Notes system. By analyzing the media keys returned by the X Pro API, we find that more than half of all requested posts contain media (65.4%).

3: We use GPT to annotate the 200 sampled posts twice and assess the consistency of its outputs across the two rounds. The outputs show strong agreement across all three variables: claims (Cohen's $\kappa = 0.920$), opinions (Cohen's $\kappa = 0.828$), and misleadingness (Cohen's $\kappa = 0.816$). In addition, we further validate the reliability of GPT-based approach using expert fact-checks from PolitiFact (*Accuracy* = 0.801, *Weighted F1* = 0.857; see details in Suppl. S9.3).

9.3.3 Characteristics of Post Authors

Built-in characteristics of post authors. We analyze several built-in features of the post authors obtained from the X Pro API. The dataset includes in total 22,915 unique post accounts who authored the requested posts. Each author account is categorized as either verified or unverified. Verified accounts are categorized as *Blue*, *Government*, or *Business*: Blue accounts correspond to individual X Premium subscribers, while government and business accounts are officially verified by X. We find that the requested posts are predominantly authored by blue accounts (72.5%), followed by unverified (13.8%), government (7.2%), and business (6.6%) accounts. In addition, we consider the number of followers (mean of 6,828,060; median of 326,732), the number of followees (mean of 1,334,676; median of 1180), and the account age, defined as the number of days from the account creation to the creation of the corresponding post. The average of the account ages in the requested posts is 3156 days, *i.e.*, ~ 9 years. These features provide insights into the authors' influence, reach, and longevity on the platform.

Misinformation exposure score and partisan score. Previous work developed a method to calculate users' exposure to misinformation from

political elites on $\mathbb{X}$ [17]. They assigned each elite (public figures and organizations) a “falsity score” based on the veracity of their statements, as determined by professional fact-checkers like PolitiFact. Users’ misinformation exposure scores were then computed by averaging the falsity scores of the elites they followed on $\mathbb{X}$. This approach allows researchers to assess the relationship between users’ exposure to elite misinformation and their sharing behaviors, as well as their estimated political ideologies. We estimate the political partisanship and misinformation exposure of the relevant post authors on $\mathbb{X}$ using the API service provided by this study. Partisanship scores range from -1 (Democrat) to 1 (Republican) and are based on the number of Democratic and Republican public figures followed by each user. The misinformation exposure score, ranging from 0 to 1 , reflects the proportion of followed public figures rated as false by PolitiFact. Out of 22,915 post authors, we successfully collect misinformation exposure scores and partisan scores for 10,492 (45.8%) authors. These scores, together with other post and author characteristics, provide key context for analyzing which posts are more likely to attract community notes and how authors’ influence and ideological orientations affect note generation. Summary statistics for the characteristics related to posts and their authors are reported in Table 9.1.

9.3.4 External Domain References in Community Notes

We extract the external domains cited in the community notes based on string matching. To assess the quality and political bias of external domains referenced in the community notes, we rely on domain ratings compiled by Lin et al. [411]. This study aggregated six sets of expert news domain ratings using an ensemble approach that combined imputation and principal component analysis, resulting in a comprehensive set of quality scores for 11,520 domains. The aggregated ratings provide a reliable measure of news quality, allowing us to evaluate the credibility of the information cited in community notes. In addition, we assess the political bias of external domains using Media Bias/Fact Check, a widely used online resource covering over 10,000 media sources, journalists, politicians, and countries [288]. Analogous to previous research [288, 412], we map bias categories to scores ranging from -1 to 1 . Specifically, the “Least Biased” and “Pro-Science” are coded as 0 . The “Left” and “Right” are coded as -1 and 1 , respectively. The “Left-Center” and “Right-Center” are coded as -0.5 and 0.5 , respectively. Finally, we successfully assign domain quality scores and domain bias scores to external domains in 367,378 community notes, accounting for 20.6% of the total notes. Together, these measures allow us to evaluate both the reliability and ideological leaning of the information cited in the community notes.

9.3.5 Identification of Top Writers

Since only top note writers can view request alerts once the associated posts reach the request threshold [405], it is essential to identify these top writers in order to evaluate whether requests actually drive note writing and coverage. The dataset, however, does not directly indicate top-writer status. To address this, we approximate the top writers using the criteria provided on the official Community Notes website: contributors must have at least 4% of their notes rated “Helpful” [413]. Based on this criterion, we identify 56,113 top note writers, representing 22.2% of all note writers (433,936). Notably, the recognition of top-writer status

also requires a writing impact score of 10 or higher, but this metric is not publicly available. Despite this limitation, our approach offers a conservative estimate that avoids underestimating the potential effect of requests on note generation.

9.3.6 Replication of Note Selection Algorithm

The Community Notes program employs a matrix-factorization-based approach to identify annotations that resonate across heterogeneous user groups [67]. We replicate this note selection algorithm to evaluate both the helpfulness and polarization of community notes [414]. The algorithm produces two key estimates: the *note intercept*, which captures overall helpfulness indicating how broadly a note is judged to be useful across raters; and the *note factor*, which, by contrast, measures polarization reflecting the extent to which contributor ratings diverge. Together, these metrics enable us to assess not only the quality of community-generated fact-checks but also the degree of consensus surrounding them. To validate our replication, we compare the note statuses generated by the reproduced algorithm with the actual production statuses. Each community note can receive one of the three statuses: Currently Rated Helpful, Needs More Ratings, and Currently Rated Not Helpful. Overall, 99.7% of replicated statuses match their production counterparts, demonstrating the reliability of our reproduction.

9.3.7 Estimation Models

In this study, we employ a series of regression models and conduct empirical analyses using our observational dataset and pre-defined variables to address the research questions.

For RQ1 and RQ2, we specify a logistic regression model to examine how contributors select posts for fact-checking:

$$\text{logit}(P(\text{Noted}_i)) = \beta_0 + \beta_1' X_i^{\text{Post}} + \beta_2' X_i^{\text{Poster}}, \quad (9.1)$$

where $P(\text{Noted}_i)$ indicates the probability of the post i receiving a community note, and β_0 is the intercept. The vector X_i^{Post} contains variables describing characteristics of post content, including sentiments, topics, content types, misleadingness, and media. Additionally, the word count of post i is included to control for the post length [74, 98]. Topics and media are represented as dummy variables, while all other post-level features are treated as continuous variables. The vector X_i^{Poster} contains variables describing characteristics of post authors, including account types, account age, followers, followees, misinformation exposure scores, and partisan scores. Account types are encoded as dummy variables, while all other author-level features are treated as continuous variables. The vectors β_1 and β_2 represent the coefficients of the post-level and author-level variables, respectively, and quantify their effects on the likelihood of receiving community notes. Since logistic regression coefficients are expressed in log-odds, we exponentiate them into odds ratios to facilitate interpretation and address our RQ1 and RQ2.

For RQ3, we specify a linear regression model to examine how the extracted features of community notes, posts, and post authors relate to

the helpfulness and polarization of community notes:

$$Y_j = \beta_0 + \beta'_1 X_j^{Note} + \beta'_2 X_j^{Post} + \beta'_3 X_j^{Poster} + \epsilon_j, \quad (9.2)$$

where Y_j represents either the helpfulness score or the polarization score of note j . β_0 is the intercept, and ϵ_j is the residual term for the linear regression. The vector X_j^{Note} includes note-specific characteristics, including domain quality and domain unbiasedness. They are treated as continuous variables. The vectors X_j^{Post} and X_j^{Poster} contains post-level and author-level features, respectively, and remain the same as defined in Eq. 9.1. Coefficient estimates for β_1 , β_2 , and β_3 reflect how note-, post-, and author-level attributes are associated with variations in note helpfulness and polarization, thus allowing us to assess RQ3.

9.4 Empirical Results

Here, we report our empirical results, structured around the three research questions. We begin by examining what types of requested posts are more likely to receive community notes (RQ1), then access the extent to which requests foster note writing (RQ2), and finally compare the quality of request-fostered notes with other notes in terms of helpfulness and polarization (RQ3).

9.4.1 Likelihood of Receiving Community Notes

Of all the requested posts, 52,862 (53.6%) received at least one community note at the time of data collection. To investigate the factors associated with this outcome (RQ1), we estimate a logistic regression model predicting the likelihood that a requested post receives a community note. The estimation results are presented in Fig. 9.2.

Post content characteristics. We analyze how post content characteristics shape the likelihood of receiving community notes. For sentiments, we find that posts expressing stronger negative sentiment are significantly less likely to receive community notes (*coef.* = -0.095 , *odds ratio* = 0.909 , $p < 0.01$). For instance, posts classified as purely negative exhibit 9.1% lower odds of receiving notes compared with non-negative posts. For topics, we find that posts related to Politics (*coef.* = -0.337 , *odds ratio* = 0.714 , $p < 0.001$) and Science & Technology (*coef.* = -0.116 , *odds ratio* = 0.890 , $p < 0.001$) are less likely to receive community notes compared to posts not associated with these topics. For example, political posts have 28.6% lower odds of receiving community notes compared to non-political posts. In contrast, posts related to Entertainment (*coef.* = 0.084 , *odds ratio* = 1.088 , $p < 0.001$) and Finance & Business (*coef.* = 0.072 , *odds ratio* = 1.075 , $p < 0.001$) are more likely to receive community notes compared to posts not associated with these topics. Additionally, posts containing media are substantially more likely to receive community notes compared to text-only posts (*coef.* = 0.205 , *odds ratio* = 1.228 , $p < 0.001$), corresponding to a 22.8% increase in the odds of receiving community notes. For claims vs. opinions, we find that posts that contain claims are associated with 14.5% higher odds of receiving a community note than posts without claims (*coef.* = 0.135 , *odds ratio* = 1.145 , $p < 0.001$), whereas posts expressing opinions are not significantly associated with the likelihood of receiving notes (*coef.* = -0.034 , *odds ratio* = 0.966 ,

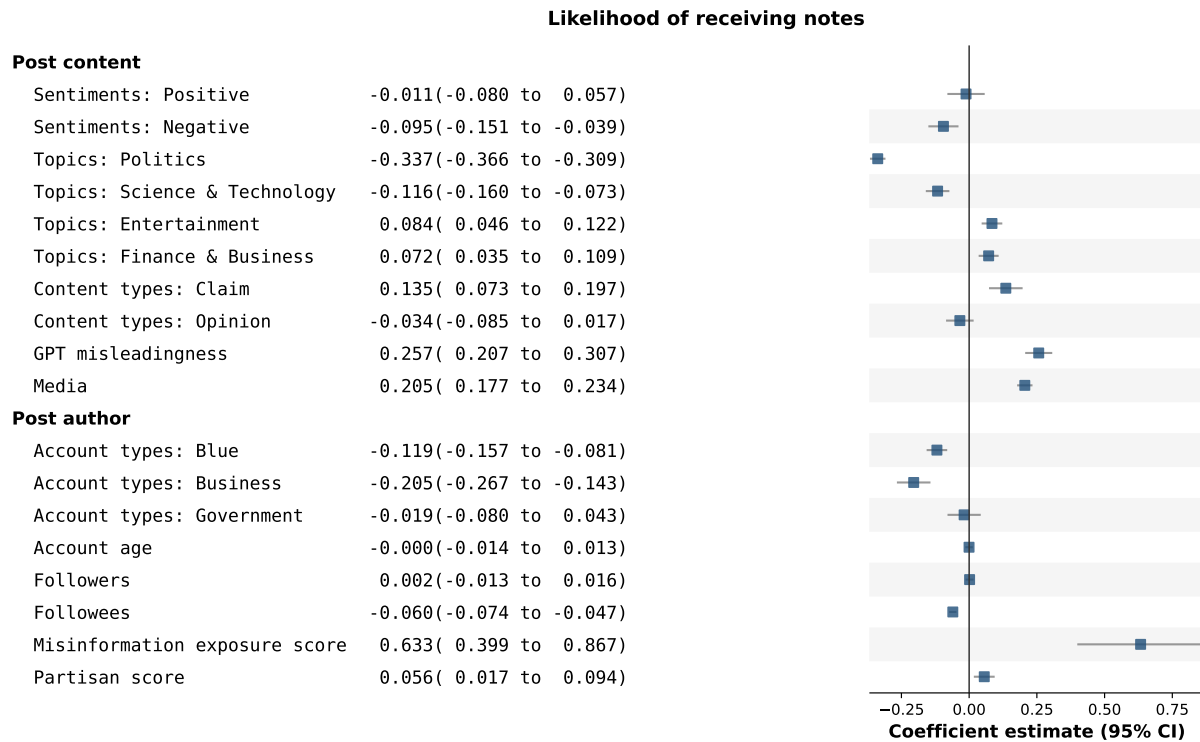


Figure 9.2: The estimation results for the logistic regression model predicting the likelihood that a requested post receives a community note. Shown are coefficient estimates with 95% Confidence Intervals (CIs). The coefficients for misinformation exposure score and partisan score are estimated based on the subset of posts containing the corresponding information. The number of words in each post is controlled during estimation but omitted in visualization for better readability. Continuous independent variables—word count, account age, number of followers, and number of followees—are z-standardized before estimation to facilitate interpretation.

$p = 0.186$). Importantly, we find that posts with higher GPT-estimated misleadingness are significantly more likely to receive community notes ($coef. = 0.257$, $odds\ ratio = 1.293$, $p < 0.001$). Specifically, posts identified by GPT as purely misleading have 29.3% higher odds of receiving community notes than non-misleading posts. This suggests that Community Notes contributors tend to prioritize posts with a higher potential for spreading misinformation.

Post author characteristics. We further evaluate how post author characteristics shape the likelihood of receiving community notes. For verified types of post authors, we find that posts from blue ($coef. = -0.119$, $odds\ ratio = 0.888$, $p < 0.001$) and business ($coef. = -0.205$, $odds\ ratio = 0.815$, $p < 0.001$) accounts have 11.2% and 18.5% lower odds of receiving community notes than posts from unverified accounts, respectively. Additionally, posts from accounts with a higher number of followees are less likely to receive community notes, with one standard deviation increase in followers corresponding to 5.8% decrease in the odds of receiving notes ($coef. = -0.060$, $odds\ ratio = 0.942$, $p < 0.001$). In contrast, misinformation exposure scores are positively associated with the likelihood of receiving community notes: posts from accounts with a misinformation exposure score of 1 have 88.3% higher odds than those from accounts with a score of 0 ($coef. = 0.633$, $odds\ ratio = 1.883$, $p < 0.001$). Moreover, posts from right-leaning accounts with a partisan score of 1 have 11.4% higher odds of being noted compared to those from left-leaning accounts with a partisan score of -1 ($coef. = 0.056$, $odds\ ratio = 1.057$, $p < 0.01$). These findings suggest that contributors prioritize posts from accounts with higher risk of spreading misinformation and right-leaning partisanship.

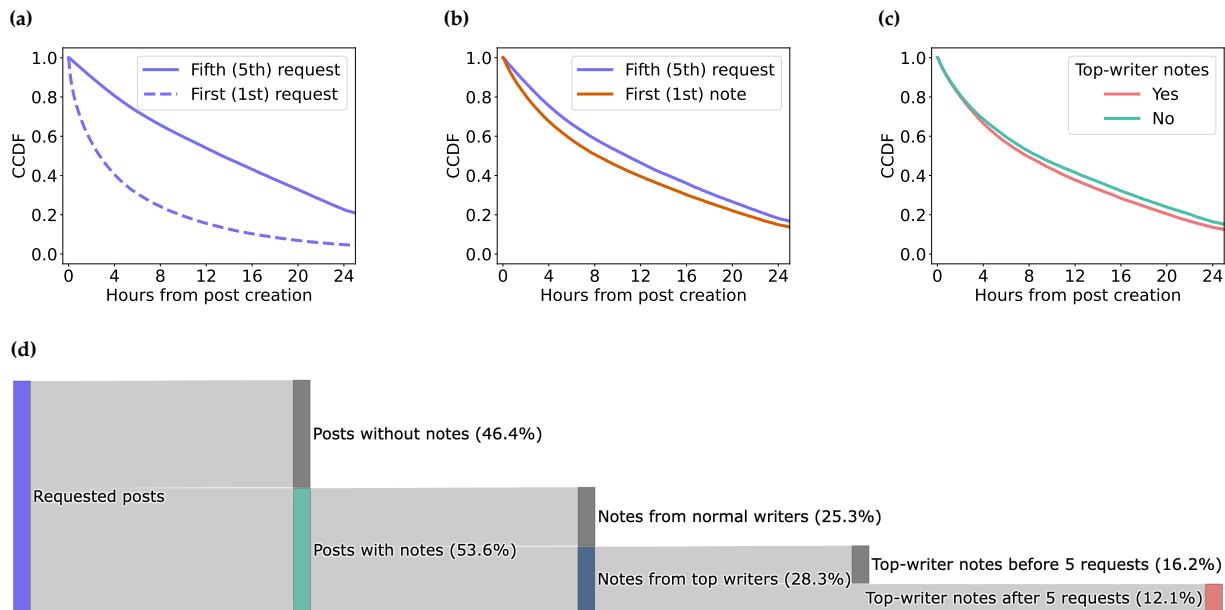


Figure 9.3: The statistics for the request timing relative note writing. (a) The Complementary Cumulative Distribution Functions (CCDFs) for the hours from post creation to the submission of first request and to the submission of fifth request. (b) The CCDFs for the hours from post creation to the submission of fifth request and to the generation of first community note, respectively. (c) The CCDFs for the hours from post creation to the notes written by top writers versus other contributors. (d) The sankey plot illustrating the estimated proportion of posts for which community notes are likely fostered by requests.

In summary, the likelihood that a requested post receives a community note is shaped by both posts' and their authors' characteristics. Requested posts with negative sentiment or authored by blue/business accounts are less likely to attract notes, as are those related to politics or science. In contrast, posts related to entertainment or finance, those containing factual claims and media, and posts flagged by GPT as more misleading are significantly more likely to receive notes. At the author level, posts from accounts with a higher misinformation exposure or right-leaning partisanship (vs. left-leaning) are also more likely to be annotated. Notably, given the heightened concerns surrounding political misinformation and the fact that political posts constitute the largest part of our dataset, we additionally restrict our analysis to political posts and find that the observed patterns remain robust (see estimation results in Suppl. S9.4). Together, our findings highlight that Community Notes contributors who write notes and requestors who submit requests may have different selection patterns that are associated with both post features and author attributes.

9.4.2 Request Timing and Request-Fostered Notes

The request feature allows non-contributors to support the Community Notes program by flagging posts, with the goal of fostering note writing. However, receiving community notes is not necessarily a direct result of the submitted requests, as contributors can also independently select and fact-check posts regardless of requests. To assess whether the request feature meaningfully fosters note creation (RQ2), we examine the timing of requests relative to note creation.

Request timing relative to note creation. As shown in Fig. 9.3a, the median time from post creation to the submission of first request is

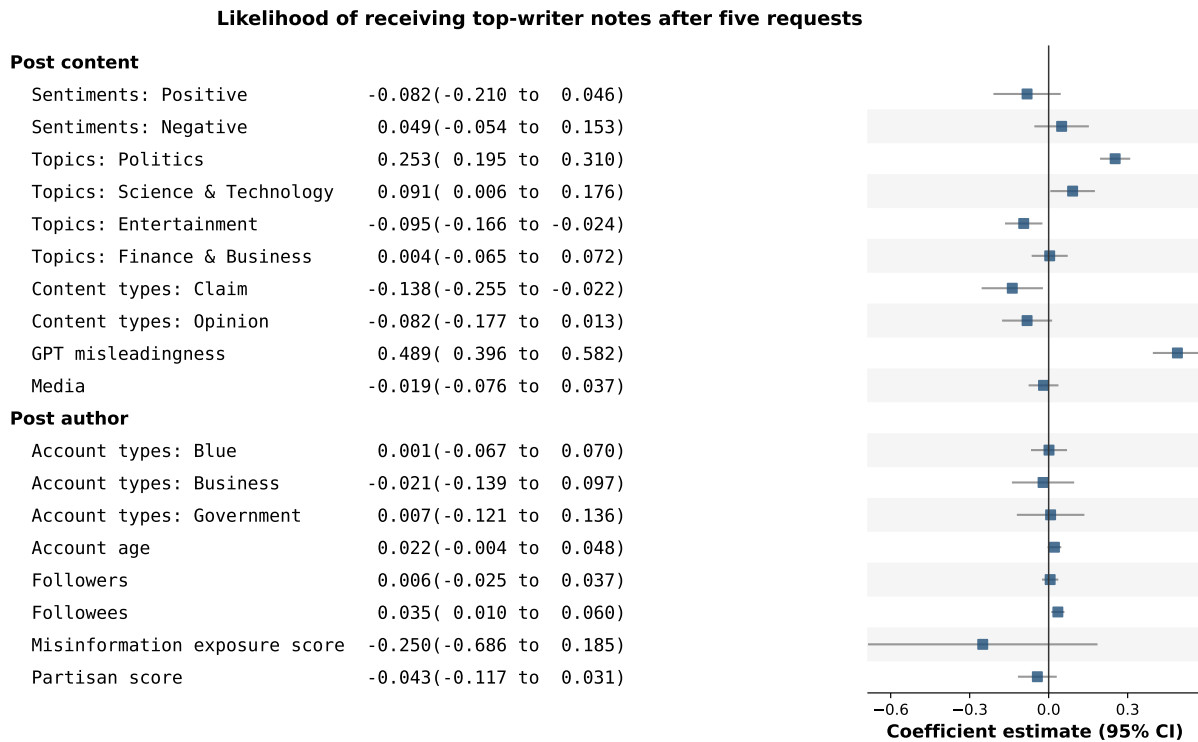


Figure 9.4: The estimation results for the logistic regression model predicting the likelihood that a requested post receives a top-writer note after the request threshold, *i.e.*, request-fostered notes. Shown are coefficient estimates with 95% CIs. The coefficients for misinformation exposure score and partisan score are estimated based on the subset of posts containing the corresponding information. The number of words in each post is controlled during estimation but omitted in visualization for better readability. Continuous independent variables—word count, account age, number of followers, and number of followees—are z-standardized before estimation to facilitate interpretation.

2.72 hours, while it takes significantly longer for a post to accumulate five requests (median of 13.40 hours; $p_{MWU} < 0.001$). We find that the response time of community notes (median of 8.23 hours) is significantly shorter than the time it typically takes for posts to receive five requests (Fig. 9.3b; $p_{MWU} < 0.001$). Furthermore, Fig. 9.3c shows that top writers (median of 7.78 hours) generate notes almost an hour faster than other contributors (median of 8.76 hours; $p_{MWU} < 0.001$). Taken together, these findings suggest that community notes tend to be written before the fifth request is submitted (*i.e.*, request threshold)—implying that requests may not be the primary driver of note creation. Additionally, because community notes often appear more quickly than request alerts, and because top writers tend to respond faster than other contributors, notes generated by top writers after the fifth request to the corresponding posts are highly likely to have been shaped by request alerts. Therefore, we consider top-writer notes generated after the request threshold as request-fostered notes. Consistent with the guidelines of the request feature, we find that only an estimated 12.1% of requested posts were likely influenced by the request activity, as their associated notes were written by top writers after the request threshold was reached (Fig. 9.3d). Therefore, the effect of the request feature in driving note creation is still limited.

Request-fostered notes from top writers. We further examine whether the request function influences the fact-checking behavior of top writers. To this end, we employ the logistic regression model as specified in Eq. 9.1 and compare top writers' selection patterns when exposed to request alerts versus when writing notes without such alerts. Here, the dependent

variable is the probability that a post receives a request-fostered note. Under the null hypothesis that requests do not affect selection behavior of top writers, the coefficient estimates for all independent variables should be statistically indistinguishable from zero; statistically significant coefficient estimates, by contrast, indicate systematic differences in post selection following request exposure. Importantly, to ensure a fair comparison, we restrict our analysis to posts on which notes were generated only by topic writers before and after the request threshold. The estimation results are presented in Fig. 9.4.

We find that top-writer notes on posts related to Politics (*coef.* = 0.253, *odds ratio* = 1.287, $p < 0.001$) and Science & Technology (*coef.* = 0.091, *odds ratio* = 1.095, $p < 0.05$) have 28.7% and 9.5% higher odds of being request-fostered, respectively, compared to posts without the corresponding topics. This suggests that top writers select more political and scientific posts when receiving request alerts. In contrast, top-writer notes on posts related to Entertainment (*coef.* = -0.095, *odds ratio* = 0.910, $p < 0.01$) and containing claims (*coef.* = -0.138, *odds ratio* = 0.871, $p < 0.05$) have 9% and 12.9% lower odds of being request-fostered, respectively. This indicates that top writers choose less entertainment-related and claim-oriented content after receiving request alerts, compared to without receiving alerts. Together, these results suggest that requests can shift contributors' attention toward political and scientific content, while diverting attention away from claim-oriented posts or those related to entertainment. In particular, the coefficient estimate for GPT-estimated misleadingness is positive and substantially significant (*coef.* = 0.489, *odds ratio* = 1.631, $p < 0.001$). Posts flagged by GPT as misleading (score = 1) have 63.1% higher odds of receiving request-fostered top-writer notes than non-misleading posts (score = 0). This suggests that top writers prioritize posts with higher potential misleadingness when guided by request alerts, compared to when no alerts are present.

In summary, out of all requested posts eligible for community evaluation, only 12.1% appear to have been generated by top writers through request alerts. Furthermore, requests typically arrive later than community notes themselves, as the time from post creation to request alerts is significantly longer than the time from post creation to note generation, limiting the request feature's potential to foster note writing. Nevertheless, by comparing notes written by top writers before and after the request threshold, we find that requests shift contributors' attention toward political content while prioritizing posts with higher potential misleadingness.⁴ Thus, although the overall effect of requests on note creation remains modest, the feature shows promise in channeling contributor attention toward high-stakes content at greater risk of misinformation.

4: To further reduce unobserved confounding at the contributor level, including heterogeneity among top writers, we additionally restrict the analysis to posts selected by the same set of top writers before and after the request threshold (see details in Suppl. S9.5). This conservative approach confirms that our main findings remain robust, particularly with respect to political content and posts with high potential misleadingness.

9.4.3 Algorithmic Evaluation of Community Notes

To address RQ3, we downloaded and rerun the source code of the note selection algorithm released by X. Fig. 9.5a presents the distribution of note intercepts and note factors from the note selection algorithm. The note intercepts reflect the perceived helpfulness of community notes across raters, with higher intercepts indicating higher helpfulness. The note factors capture the polarization in ratings, with values deviating from 0 indicating greater disagreement among raters regarding the helpfulness of community notes. Here, we categorize all community notes into three groups: *writer-only* (notes on posts that received fewer than five requests and were solely written by contributors), *request-fostered* (notes authored

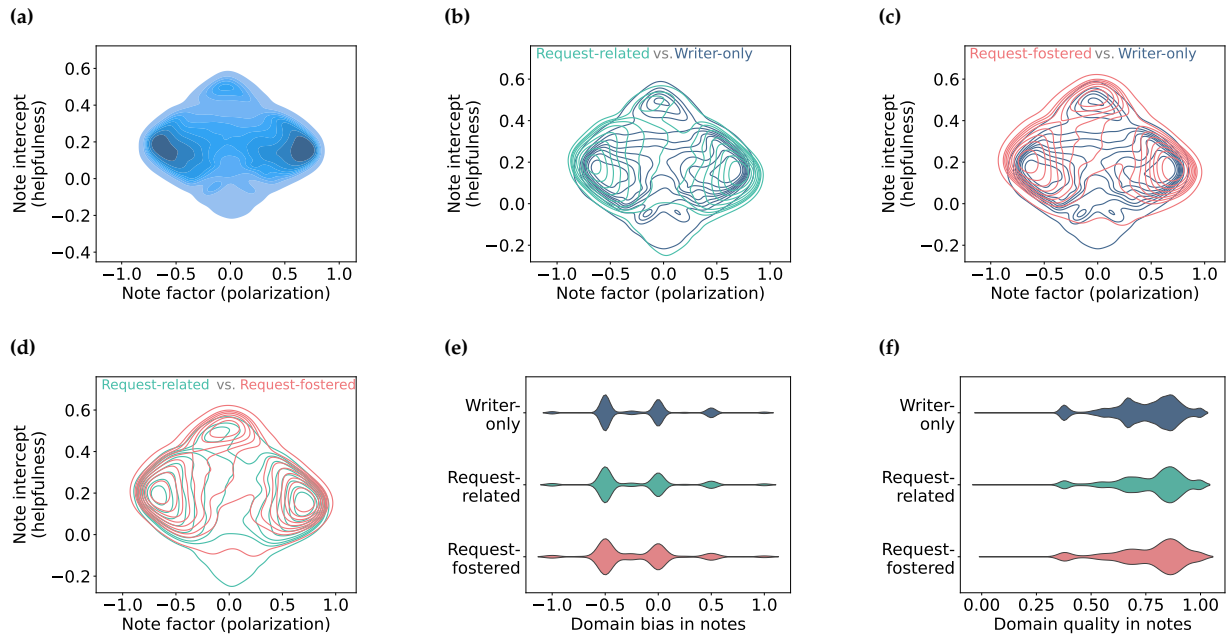


Figure 9.5: Overview of note evaluations and source domains in community notes. (a) The distribution of estimated note intercepts (helpfulness) and estimated note factors (polarization) from the note selection algorithm. (b) The distributions of estimated note intercepts (helpfulness) and note factors (polarization) from the note selection algorithm between request-related notes and writer-only notes. (c) The distributions of estimated note intercepts (helpfulness) and note factors (polarization) from the note selection algorithm between request-fostered notes and writer-only notes. (d) The distributions of estimated note intercepts (helpfulness) and note factors (polarization) from the note selection algorithm between request-related notes and request-fostered notes. (e) The violin plots showing the distributions of domain bias in community notes across the three categories: writer-only, request-related, and request-fostered. (f) The violin plots showing the distributions of domain quality in community notes across the three categories: writer-only, request-related, and request-fostered.

by top writers after the fifth request, likely influenced by requests), and *request-related* (all remaining notes on posts that received five or more requests but not classified as “request-fostered”). Notably, request-related notes may also be written by top writers; however, unlike request-fostered notes, they were produced before the threshold of request alerts. We then analyze the helpfulness and polarization (measured as the absolute value of note factors) of community notes across these categories.

Helpfulness and polarization of request-fostered notes. We examine the distributions of note intercepts and note factors across writer-only notes, request-related notes, and request-fostered notes.

- Request-related notes vs. writer-only notes (Fig. 9.5b): Request-related notes exhibit lower helpfulness (mean = 0.172 vs. 0.178; $p_{MWU} < 0.001$) and greater polarization (mean = 0.453 vs. 0.343; $p_{MWU} < 0.001$), compared to writer-only notes.
- Request-fostered notes vs. writer-only notes (Fig. 9.5c): Request-fostered notes are rated as more helpful (mean = 0.231 vs. 0.178; $p_{MWU} < 0.001$) but also more polarized (mean = 0.424 vs. 0.343; $p_{MWU} < 0.001$) than writer-only notes.
- Request-related notes vs. request-fostered notes (Fig. 9.5d): Request-fostered notes are rated as more helpful (mean = 0.231 vs. 0.172; $p_{MWU} < 0.001$) and less polarized (mean = 0.424 vs. 0.453; $p_{MWU} < 0.001$) than request-related notes.

Taken together, these findings indicate that, although requests in general are associated with less helpful and more polarized notes, request-fostered notes written by top writers tend to achieve higher quality—

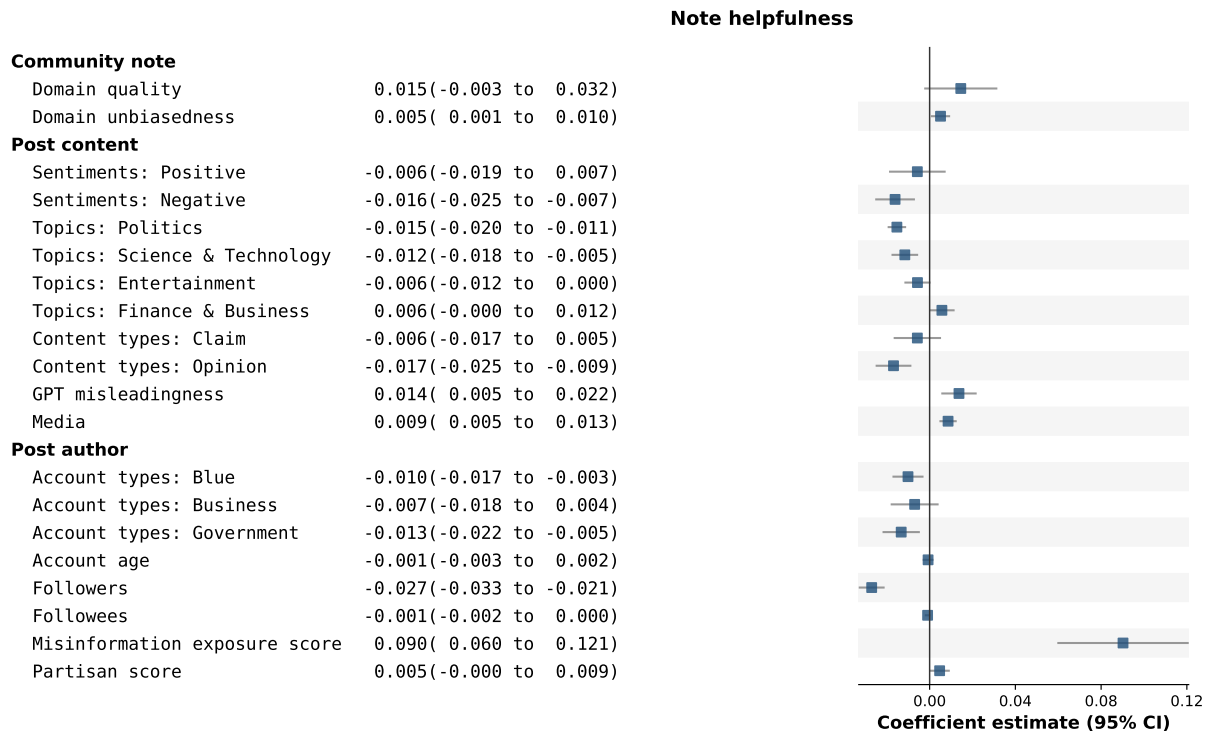


Figure 9.6: The estimation results for the linear regression model predicting note helpfulness. Shown are coefficient estimates with 95% CIs. The number of words in each post is controlled during estimation but omitted in visualization for better readability. Continuous independent variables—word count, account age, number of followers, and number of followees—are z-standardized before estimation to facilitate interpretation.

being both more helpful and less polarized than request-related notes. To further substantiate this pattern, we estimate a logistic regression predicting whether a note is request-fostered versus request-related using helpfulness and polarization as predictors. The results show a strong positive association for helpfulness ($coef. = 2.528$, $odds\ ratio = 12.530$, $p < 0.001$) and a significant negative association for polarization ($coef. = -0.231$, $odds\ ratio = 0.794$, $p < 0.001$), confirming that request-fostered notes stand out within the request context by being both more helpful and less polarized than other notes. Subsequently, we investigate what factors could explain this quality advantage.

Since contributors are required to cite external sources when writing community notes, we evaluate cited domains in terms of political bias and information quality to probe possible explanations for the higher quality of request-fostered notes than other notes. Overall, we find that community notes tended to cite more left-leaning domains than right-leaning ones (Fig. 9.5e). Domains in request-related (mean of -0.210) are slightly more left-leaning than those in writer-only notes (mean of -0.185 ; $p_{MWU} < 0.001$). The political bias of domains in request-fostered notes (mean of -0.214) has no statistically significant difference from request-related notes ($p_{MWU} = 0.352$). In terms of information quality (Fig. 9.5f), request-related notes cite higher-quality domains (mean of 0.769) than writer-only notes (mean of 0.742 ; $p_{MWU} < 0.001$), whereas request-fostered notes (mean of 0.767) do not differ from request-related notes ($p_{MWU} = 0.834$). These findings suggest that the higher quality of request-fostered notes is unlikely to be explained by domain choices alone.

Because Community Notes contributors are anonymized within the system, rater bias and polarization toward specific writers' identities

(*e.g.*, partisanship) are reduced [66]. Consequently, observed differences in helpfulness or polarization are more likely to reflect the types of posts and authors contributors choose to fact-check. We therefore examine which notes, posts, and authors are associated with higher helpfulness or polarization.

Factors associated with note helpfulness. The estimation results from the linear regression model for note helpfulness (as specified in Eq. 9.2) are shown in Fig. 9.6. We find that the unbiasedness (*i.e.*, neither left-leaning nor right-leaning) of external domains cited in community notes is positively associated with note helpfulness ($coef. = 0.005, p < 0.05$), suggesting that notes referencing politically neutral sources are perceived as more helpful compared to those citing left- or right-leaning domains. In contrast, the information quality of external domains is not significantly associated with helpfulness ($coef. = 0.015, p = 0.095$).

Regarding post content, notes on posts with higher negative sentiment are rated as less helpful ($coef. = -0.016, p < 0.01$). Posts related to Politics ($coef. = -0.015, p < 0.001$) and Science & Technology ($coef. = -0.012, p < 0.001$) tend to receive less-helpful notes compared to posts without these topics. Additionally, posts with higher opinion scores are associated with lower note helpfulness ($coef. = -0.017, p < 0.001$). Conversely, community notes on posts containing media ($coef. = 0.009, p < 0.001$) or with higher estimated misleadingness ($coef. = 0.014, p < 0.01$) are rated as more helpful. With respect to post authors, notes on posts from the blue ($coef. = -0.010, p < 0.01$) and government ($coef. = -0.013, p < 0.01$) accounts, as well as from authors with more followers ($coef. = -0.027, p < 0.001$), are less helpful compared to those from unverified accounts with fewer followers. In contrast, higher misinformation exposure scores of post authors are positively associated with note helpfulness ($coef. = 0.090, p < 0.001$).

In summary, the helpfulness of community notes is positively associated with (i) citations to unbiased domains, (ii) posts containing media, (iii) post misleadingness, and (iv) misinformation exposure scores of post authors. Conversely, the helpfulness of community notes is negatively associated with (i) negative sentiment in posts, (ii) political and scientific topics, (iii) opinionated content, (iv) blue and government accounts, and (v) the number of followers.

Factors associated with note polarization. The estimation results for a linear regression model predicting note polarization are shown in Fig. 9.7. Community notes that cite politically unbiased domains tend to have reduced polarization across their ratings ($coef. = -0.019, p < 0.001$), whereas citations to domains with higher information quality are associated with increased polarization ($coef. = 0.046, p < 0.01$).

In terms of post content, positive sentiment in posts is negatively associated with note polarization ($coef. = -0.038, p < 0.001$), but negative sentiment shows the opposite pattern ($coef. = 0.017, p < 0.05$). Posts related to Politics are linked to higher polarization ($coef. = 0.054, p < 0.001$), whereas posts related to Finance & Business ($coef. = -0.022, p < 0.001$) are associated with lower polarization. Community notes on posts containing stronger claims ($coef. = 0.025, p < 0.001$) or more opinionated content ($coef. = 0.041, p < 0.001$) tend to be more polarized. In contrast, community notes on posts that include media ($coef. = -0.027, p < 0.001$) or exhibit higher misleadingness ($coef. = -0.032, p < 0.001$) have lower polarization scores. For post authors, posts from all types of verified accounts—blue ($coef. = 0.044, p < 0.001$), business ($coef. = 0.038, p < 0.001$), and government ($coef. = 0.054, p < 0.001$)—tend to receive

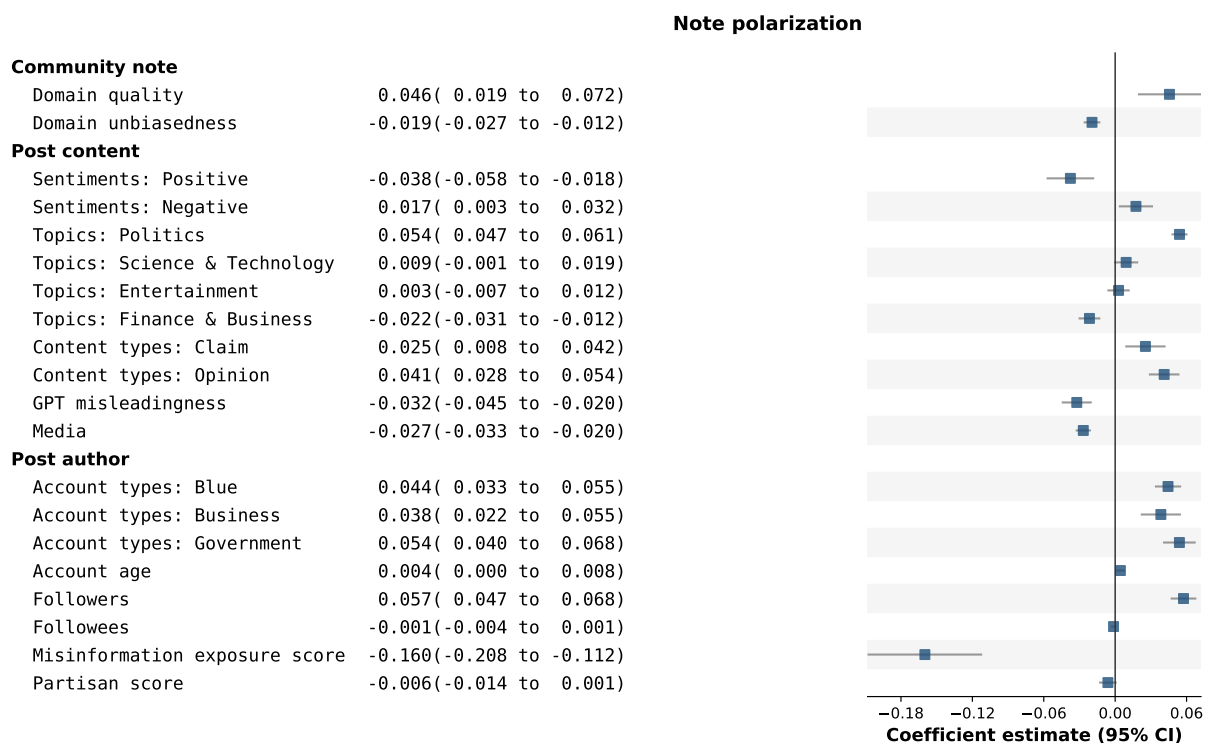


Figure 9.7: The estimation results for the linear regression model predicting note polarization. Shown are coefficient estimates with 95% CIs. The number of words in each post is controlled during estimation but omitted in visualization for better readability. Continuous independent variables—word count, account age, number of followers, and number of followees—are z-standardized before estimation to facilitate interpretation.

community notes with higher polarization compared to posts from unverified accounts. Additionally, posts authored by accounts with more followers are associated with greater note polarization (*coef.* = 0.057, $p < 0.001$), and posts from accounts with higher misinformation exposure scores are more likely to receive community notes with reduced polarization (*coef.* = -0.160, $p < 0.001$).

In summary, the polarization of community notes is positively associated with (i) high quality of external domains, (ii) negative sentiment, (iii) political topics, (iv) posts containing claims or opinionated content, and (v) verified accounts with many followers. Conversely, the polarization of community notes is negatively associated with (i) citations to unbiased domains, (ii) positive sentiment, (iii) posts containing media, (iv) misleadingness of post content and (v) misinformation exposure scores of post authors.

Notably, through the analysis of both note helpfulness and polarization, we observe that post misleadingness is positively associated with note helpfulness but negatively associated with note polarization, whereas political content exhibits the opposite pattern. Together with our findings in Section 9.4.2, which show that requests shift top writers' attention toward political content while prioritizing posts with higher potential misleadingness, our analysis provides additional insight: the higher helpfulness and lower polarization of request-fostered notes can be partially interpreted as reflecting top writers' selective fact-checking of misleading posts.

9.4.4 Summary of Main Findings

In this study, based on the request feature implemented on X and associated requests, posts, and community notes, we conduct a comprehensive analysis on the Community Notes system with respect to content selection, contributor behavior, and note evaluation. Overall, our analysis highlights three key findings.

- Requested posts are more likely to receive notes when they are associated with entertainment, finance, media, claims, high misleadingness, or authors with high misinformation exposure, but less so for negative, political or scientific content. This suggests that Community Notes contributors and requestors have distinct selection patterns.
- Only an estimated 12.1% of the requested posts appear to receive request-fostered notes from top writers, with many notes written independently of the request feature. Nevertheless, requests can shape contributors' behavior by directing their focus toward politically salient content, while still prioritizing posts with higher risk of misinformation.
- Request-fostered notes exhibit higher helpfulness and lower polarization compared to notes generated prior to the request threshold or by non-top writers, a pattern that may partly reflect top writers selectively fact-checking posts with higher misleadingness.

9.5 Discussion

Although the Community Notes system has emerged as a promising model of crowdsourced fact-checking, its speed and coverage remain limited, constraining its scalability on social media platforms [76, 398]. To address these challenges, X recently introduced the Request Community Note feature, which opens a new channel for fact-checking participation and enables a broader set of users to actively solicit notes on specific posts. In this study, we analyze a large-scale dataset of 98,685 posts surfaced through the request feature, along with their associated notes, to examine how this feature influences X's Community Notes system. Our findings shed light on the role of requests in shaping what content is fact-checked, how contributors engage, and the quality of resulting notes.

9.5.1 Research Implications

We find that more than half of the requested posts (53.6%) ultimately receive community notes. Yet, the processes of submitting requests and generating notes are largely independent, with only 12.1% of posts receiving annotations that can be potentially attributed to the request feature. This suggests that requests do not strongly determine contributor behavior. Instead, the relative independence of the two processes makes requests a useful lens for examining how Community Notes contributors select posts for fact-checking and how their choices align—or diverge—from the priorities expressed by users who submit requests (*i.e.*, requestors).

Content selection by contributors. Our results show that contributors' decisions to fact-check requested posts are shaped by both post content and author characteristics. Specifically, contributors are less likely to write notes for posts with negative sentiment, authored by blue or business

5: To validate our findings on content selection by contributors and rule out confounding from request-driven activity, we exclude posts with request-fostered community notes (as identified in Section 9.4.2) and repeat our analysis. The results remain robust (see details in Suppl. S9.6).

accounts, or focused on politics or science. In contrast, posts related to entertainment or finance, those containing claims or media, and those flagged by GPT as more misleading are significantly more likely to receive notes. At the post author level, posts from accounts with higher misinformation exposure or right-leaning partisanship are also more likely to be annotated. The misalignment between the types of posts surfaced by requestors (*e.g.*, political content) and those contributors prioritize for annotation underscores a divergence in fact-checking selection.⁵

On the one hand, Community Notes contributors tend to fact-check claim posts with higher estimated misleadingness and from authors with higher misinformation exposure scores. This suggests that, compared to requestors, contributors adopt stricter standards, prioritizing content with greater potential for misinformation and harm. On the other hand, previous research finds that contributors often fact-check posts that are relatively straightforward to verify or contain claims already addressed in earlier expert fact-checks [386]. Similarly, LLMs perform strongly when processing logically structured claims but struggle with more complex ones [62]. This suggests that, although contributors appear to focus their attention on posts with higher GPT-estimated misleadingness, their fact-checking may simultaneously be biased toward content that is easier to assess. Such a focus can improve efficiency and support the production of high-quality notes, but it also risks overlooking politically salient or ambiguous posts that requestors perceive as urgent. Addressing this gap may require hybrid models of collaboration, where professional fact-checkers complement contributors by targeting content that is more ambiguous, emergent, or difficult to verify [45].

Partisan selection biases between contributors and requestors. Notably, the human-centered fact-checking approaches—whether expert-based or crowdsourced—cannot fully avoid political bias in the selection of targets [73, 145, 150]. Our results reveal significant differences in how requestors and contributors approach political content and partisan cues. Previous work has shown that users preferentially challenge content authored by those with opposite partisan leanings, and Republicans are flagged more often than Democrats for sharing misinformation on X’s Community Notes [50, 378, 412]. Our findings echo this pattern: contributors are more likely than requestors to write notes on posts authored by Republicans, whereas requestors are comparatively more likely to surface posts authored by Democrats. This divergence raises an important interpretive challenge. On the one hand, it may reflect partisan selection biases, with anonymized contributors and requestors each applying their own subjective judgments when deciding what to fact-check [415]. On the other hand, it could mirror underlying asymmetries in the misinformation landscape, where different political groups are disproportionately represented in the spread of misleading content. The coexistence of such tendencies highlights the complexity of maintaining balance and fairness in decentralized fact-checking systems like Community Notes.

Response of top writers to requests. Although the direct impact of the request feature on fostering note writing and coverage remains modest, requests appear to shift the attention of top writers: they are more likely to engage with political posts and focus on content with higher estimated misleadingness. By concentrating on posts with greater misinformation potential, request-fostered notes from top writers are evaluated as more helpful and less polarized than other notes on requested posts. In this sense, contributors play a crucial role as gatekeepers, directing community fact-checking toward high-risk content.

At the same time, the request feature indirectly amplifies the influence of these contributors in shaping what gets fact-checked. A small elite of writers (22.2%) is responsible for nearly half of all community notes (49.7%), underscoring the uneven distribution of labor characteristic of volunteer-based systems. This concentration of fact-checking activity raises concerns about scalability and resilience: if the system relies disproportionately on a narrow core of expert-like participants, its long-term sustainability may be vulnerable to burnout, disengagement, or shifts in contributor incentives [45]. Moreover, while the reliance on elite contributors enhances quality and consensus, it also complicates the platform's vision of broad, community-driven participation. Rather than democratizing fact-checking, requests may consolidate authority in the hands of a few, blurring the line between peer production and expert review.

Reliability and vulnerability of the note selection algorithm. We provide a comprehensive analysis of how the helpfulness and polarization of community notes are shaped by both content and source factors. Posts with higher estimated misleadingness and authored by accounts with greater misinformation exposure are more likely to elicit helpful annotations. This finding supports the view that Community Notes can function as a reliable mechanism for surfacing accurate fact-checks and correcting misinformation [70].

However, our findings on polarization also highlight a critical vulnerability. Notes on political or opinionated content tend to provoke divided evaluations, even when they cite high-quality domains, which resonates with previous research [385]. Such polarization does not necessarily imply inaccuracy, but can reflect partisan disagreement over credibility, legitimacy of sources, or interpretive framing [377]. Therefore, the Community Notes system faces a significant challenge: factually accurate annotations may still fail to gain consensus if they are evaluated through polarized lenses. Addressing this limitation points to the need to improve the note selection algorithm to better balance factual accuracy with the practical requirement of fostering consensus among heterogeneous communities.

9.5.2 Practical Implications

Our findings provide several actionable insights for the design of community-based fact-checking systems on social media platforms, such as X's Community Notes.

Reducing the delay of request alerts. Requests currently function less as a mechanism to increase overall note volume and coverage, partly due to differences in selection patterns between contributors and requestors, and partly due to delays in reaching the request alert threshold relative to note generation. Although the first request may appear early in a post's diffusion, it often takes considerably longer to accumulate enough requests to trigger the alert. To reduce these delays, the request mechanism could be adapted to customize thresholds based on requestor reputation, granting lower thresholds to high-reputation users while maintaining safeguards for others. For example, X is already testing a system that computes helpfulness scores for requestors, such that users with higher scores require fewer co-requests before their note requests are surfaced to contributors [405].

Leveraging LLMs to mitigate contributor bias and workload. Given recent advances in LLMs, these models could be incorporated into the Community Notes system to automatically write notes for requested posts with high estimated misleadingness. Such an integration would help mitigate the potential selection bias in note writers and also alleviate their workload. For instance, X has opened the AI Note Writer API to developers, enabling the design of tools that assist in writing notes on the requested posts [416]. Beyond writing notes on posts requested by users, LLM-based tools can be expanded to provide proactive support across the platform—for example, issuing warnings to authors when they attempt to publish potentially misleading posts, reviewing posts after publication, and surfacing content with high misleadingness for community evaluation. Together, these applications would help broader and more uniform coverage of online content, complementing the request feature and strengthening the scalability of community-based fact-checking.

Addressing polarization and improving consensus. The current note selection algorithm faces a significant challenge: community notes, even when factually accurate, on posts related to politics and from high-influence users (*e.g.*, verified accounts with many followers) tend to provoke polarized evaluations and fail to reach consensus among heterogeneous communities [377]. To address this challenge, platforms could consider two directions for improvement. (i) Introducing an review layer by experts or contributors could help access highly polarized but potentially helpful notes, guiding decisions on which annotations to display [45]. In addition, the algorithm could assign higher weights to ratings from top contributors when aggregating evaluations. (ii) Our findings indicate that notes citing unbiased sources are evaluated as more helpful and less polarized, consistent with previous research [288]. Given this, platforms could provide contributor guidelines or incentives to encourage the use of neutral sources. Furthermore, LLMs can be leveraged to synthesize existing community notes, improve clarity, and then foster consensus-building across diverse user communities [398].

Taken together, these actionable insights can inform the design and enhancement of community-based fact-checking systems. In particular, LLMs have promise for transforming the Community Notes system in streamlining request processing and assisting in note generation, thereby helping the system remain both reliable and responsive to user demands [396].

9.5.3 Limitations and Future Work

Our study has several limitations and offers directions for future research. First, our analysis is based on observational data from X, which constrains causal inferences. While we examine various characteristics—related to posts, their authors, and associated notes—and identify their associations with note outcomes, we cannot definitively determine whether certain factors directly cause higher helpfulness or lower polarization. Future studies could explore causal relationships and clarify underlying mechanisms through controlled survey experiments or platform-supported interventions. Second, while our GPT-based approach provides scalable assessments of content types and potential misinformation, it remains subject to hallucinations and model-specific biases [417], and may not fully capture emerging or highly complex misleading content [62]. To mitigate these concerns, we conduct comprehensive validations using both assistant-annotated and expert-annotated datasets to demonstrate

the reliability of GPT-generated outputs for our analysis. With continued advances in LLM-based fact-checking approaches, annotation accuracy is expected to further improve. Future work is encouraged to validate our findings when more advanced models are available. Third, given that the top-writer status is not available in our dataset, we cannot perfectly distinguish request-fostered notes from contributor-driven notes. Nevertheless, our estimation method ensures that the effect of the request feature in fostering note writing is not underestimated. Finally, contributors and requestors are anonymized in the Community Notes system, preventing the analysis of individual-level behavior, motivation, or expertise. While this anonymization can reduce rater bias, it also limits understanding of how personal experience or identity may influence content selection and note quality. Future work could explore these dynamics in contexts where such information can be collected ethically through survey experiments or accessed under appropriate privacy safeguards on platforms.

Motivated by persistent societal harms caused by misinformation – especially during public health crises and political events – this thesis investigates the multifaceted challenges of misinformation on social media by examining how it spreads, how it is fact-checked, and how emerging fact-checking intervention, *i.e.*, Community Notes program, performs in real-world settings. Combining large-scale observational data, quasi-experimental methods, and rigorous analyses, this thesis provides a comprehensive framework for understanding and mitigating the spread of misinformation in the digital age.

The journey of this research follows a logical progression from understanding the misinformation problem and limitations in professional fact-checking approach (**Part I**) to evaluating (**Part II**) and improving (**Part III**) community-based fact-checking solution.

In **Part I**, this thesis contributes to a deeper understanding of how users engage with misinformation. Focusing on health crises, it demonstrates that misinformation is often multi-topic and frequently intertwined with conspiracy narratives, which increases user engagement. By distinguishing between news source sharers and post viewers, the thesis reveals that these two user groups play different roles in misinformation diffusion: sharers contribute to the persistence and repeated circulation of misleading content, while viewers are more responsive to topic diversity and conspiratorial framing. These findings highlight the importance of accounting for user heterogeneity when designing interventions to curb misinformation.

Additionally, the thesis reveals constraints in the practices of professional fact-checking organizations. By analyzing long-term data on fact-checked political statements in the U.S., it shows that false claims are more likely than true ones to mention political elites and that such mentions exhibit clear partisan and temporal patterns. These results provide empirical evidence relevant to ongoing debates about politicization, neutrality, and trust in professional fact-checking, emphasizing the need for greater transparency in claim selection and prioritization.

Part II provides the first multi-dimensional causal evaluation of X's Community Notes. Across multiple studies, this thesis demonstrates that community notes can substantially reduce the spread of misleading content and increase post deletion once annotations are displayed. At the same time, the findings reveal important limitations: community notes often appear too late to meaningfully intervene during the most viral phase of diffusion, and their overall platform-level impact is therefore constrained. In addition, the thesis shows that community fact-checks elicit strong emotional responses, including increased negativity and moral outrage, underscoring the social and emotional consequences of public correction.

In **Part III**, this thesis examines how community-based fact-checking system evolves and how its design affects robustness and scalability. It identifies vulnerabilities in consensus-based algorithms, particularly post-display polarization that can destabilize initially helpful notes. It also evaluates the role of expanded participation through user-initiated

fact-checking requests, showing both the promise of scalable user involvement and the challenges posed by partisan misalignment and selective engagement. Together, these results offer system-level insights into how community-based fact-checking can be made more resilient and effective.

Overall, this thesis makes significant contributions to misinformation research by deepening the understanding of misinformation diffusion during health crises, identifying the constraints of traditional expert-based fact-checking, and providing a rigorous evaluation of community-based interventions. Collectively, these findings provide actionable insights for social media platforms seeking to develop more reliable, timely, and resilient approaches to safeguarding information integrity. As the digital landscape continues to evolve, the shift from centralized gatekeeping to transparent, consensus-driven moderation offers a promising path toward a more informed and resilient public discourse.

SUPPLEMENTARY MATERIALS

Ethical Considerations

This research has received ethical approval from the Ethics Review Panel of the University of Luxembourg (ref. ERP 23-053 REMEDIS). All analyses are based on publicly available data or user experiments conducted with appropriate approval. The user study in Chapter 5 was additionally approved by the Institutional Review Board at Princeton University. All data collection from X complies with the platform’s developer rules. To respect privacy, we explicitly do not publish usernames in our paper and report only aggregate results.

S2

Supplementary Materials for Chapter 2

S2.1 Estimation Results for Main Analysis

The full estimation results for the analysis in the main paper are reported in Tables S2.1–S2.3. Specifically, Table S2.1 reports the estimation results for the topic diversity. Table S2.2 reports the estimation results for the engagement at news source sharer level. Table S2.3 reports the estimation results for the engagement at post viewer level. Additionally, Table S2.4 reports the estimation results for the subset analysis of post viewer engagement within false news posts.

	(1) News claim	(2) News post	(3) Claim → Post
Falsehood	0.388*** (0.036)	0.751*** (0.024)	0.659*** (0.022)
Conspiracy	0.061 (0.089)	0.136*** (0.015)	0.185*** (0.019)
Words	0.391*** (0.013)	0.373*** (0.004)	0.390*** (0.004)
Media		0.022*** (0.006)	0.035*** (0.008)
Verified		0.037*** (0.008)	0.004 (0.010)
AccountAge		0.003 (0.003)	0.001 (0.003)
Followers		−0.001 (0.003)	−0.019*** (0.004)
Followees		−0.013*** (0.002)	−0.015*** (0.003)
Emotions	✓	✓	✓
Language	✓	✓	✓
MonthYear	✓	✓	✓
Intercept	−0.078 (0.110)	−0.653*** (0.039)	−0.718*** (0.047)
News-level REs	✗	✓	✓
N	7273	70,904	70,904
R ²	0.179	✗	✗

Table S2.1: Estimation results for the topic diversity in news claims [Column (1)], the topic diversity in news posts [Column (2)], and the changes in topic diversity from news claims to associated posts [Column (3)]. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table S2.2: Estimation results for news lifetime [Column (1)] and post count [Column (2)]. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) News lifetime	(2) Post count
Falsehood	1.239*** (0.046)	2.230*** (0.083)
Conspiracy	0.452* (0.189)	0.493 (0.335)
Diversity	-0.002 (0.005)	-0.025 (0.026)
Falsehood $\times$ Diversity	-0.075 (0.039)	0.079 (0.085)
Words	-0.080*** (0.010)	-0.614*** (0.030)
Emotions	✓	✓
Language	✓	✓
MonthYear	✓	✓
Intercept	2.491*** (0.282)	2.060*** (0.198)
N	7273	7273
R^2	0.481	0.188

Table S2.3: Estimation results for repost count [Column (1)], like count [Column (2)], and reply count [Column (3)]. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) Repost count	(2) Like count	(3) Reply count
Falsehood	-0.256*** (0.063)	-0.155* (0.063)	0.400*** (0.058)
Conspiracy	0.342*** (0.065)	0.373*** (0.059)	0.365*** (0.056)
Diversity	0.181*** (0.031)	0.192*** (0.030)	0.137*** (0.030)
Falsehood $\times$ Diversity	-0.125*** (0.035)	-0.093** (0.033)	-0.091** (0.033)
Words	0.519*** (0.017)	0.467*** (0.016)	0.443*** (0.015)
Media	0.788*** (0.028)	0.596*** (0.027)	0.432*** (0.027)
Verified	3.037*** (0.035)	3.071*** (0.034)	2.367*** (0.033)
AccountAge	0.091*** (0.012)	0.084*** (0.011)	0.041*** (0.012)
Followers	0.175*** (0.015)	0.180*** (0.014)	0.217*** (0.013)
Followees	0.546*** (0.024)	0.383*** (0.022)	0.207*** (0.015)
Emotions	✓	✓	✓
Language	✓	✓	✓
MonthYear	✓	✓	✓
Intercept	0.129 (0.169)	1.055*** (0.155)	-1.398*** (0.153)
News-level REs	✓	✓	✓
N	70,904	70,904	70,904
R^2	$\times$	$\times$	$\times$

	(1) Repost count	(2) Like count	(3) Reply count
Conspiracy	0.356*** (0.075)	0.389*** (0.066)	0.378*** (0.062)
Diversity	0.073*** (0.021)	0.134*** (0.019)	0.053** (0.018)
Words	0.502*** (0.025)	0.416*** (0.022)	0.454*** (0.021)
Media	0.895*** (0.037)	0.712*** (0.034)	0.525*** (0.033)
Verified	2.624*** (0.048)	2.564*** (0.045)	1.969*** (0.042)
AccountAge	0.097*** (0.016)	0.113*** (0.014)	0.079*** (0.014)
Followers	0.502*** (0.046)	0.616*** (0.046)	0.602*** (0.036)
Followees	0.717*** (0.035)	0.512*** (0.031)	0.269*** (0.022)
Emotions	✓	✓	✓
Language	✓	✓	✓
MonthYear	✓	✓	✓
Intercept	0.129 (0.215)	1.290*** (0.190)	−0.451* (0.186)
News-level REs	✓	✓	✓
<i>N</i>	52,271	52,271	52,271
<i>R</i> ²	✗	✗	✗

Table S2.4: Estimation results for repost count [Column (1)], like count [Column (2)], and reply count [Column (3)] in the subset of false news posts. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

S2.2 Cross-Correlations & Variance Inflation Factors

We provide the cross-correlations to show how the independent variables relate to each other at the news claim and news post levels. Table S2.5 reports the cross-correlations among the variables in news claims. Table S2.6 reports the cross-correlations among the variables in news posts.

Additionally, we examine the Variance Inflation Factors (VIFs) among the independent variables to mitigate the concern about multicollinearity. The variance inflation factors are reported in Table S2.7. We find that the variance inflation factors for the independent variables are all close to one and well below the critical threshold of four, which indicates that multicollinearity is not an issue of our analysis.

Table S2.5: Correlations of variables in news claims

	Falsehood	Conspiracy	Anger	Disgust	Fear	Joy	Sadness	Surprise	Words
Falsehood	1.000								
Conspiracy	0.221	1.000							
Anger	-0.010	-0.018	1.000						
Disgust	0.094	0.014	0.182	1.000					
Fear	0.030	-0.018	0.159	-0.027	1.000				
Joy	-0.007	0.026	-0.120	-0.118	-0.216	1.000			
Sadness	0.082	-0.005	0.094	0.108	0.283	-0.185	1.000		
Surprise	-0.009	-0.020	0.036	0.050	-0.087	-0.001	0.020	1.000	
Words	-0.371	-0.063	0.020	-0.047	0.077	0.100	-0.016	0.017	1.000

Table S2.6: Correlations of variables in news posts

	Falsehood	Conspiracy	Anger	Disgust	Fear	Joy	Sadness	Surprise	Words	Media	Verified	AccountAge	Followers	Followees
Falsehood	1.000													
Conspiracy	0.151	1.000												
Anger	0.038	0.002	1.000											
Disgust	0.122	-0.019	0.139	1.000										
Fear	0.075	-0.054	0.089	-0.063	1.000									
Joy	-0.020	0.062	-0.142	-0.155	-0.254	1.000								
Sadness	0.133	-0.065	0.093	0.014	0.375	-0.220	1.000							
Surprise	0.051	-0.007	0.021	-0.060	-0.124	0.057	-0.015	1.000						
Words	0.155	0.016	0.060	-0.017	0.046	0.033	0.055	0.057	1.000					
Media	-0.076	-0.051	-0.017	-0.031	0.013	0.007	0.000	-0.033	0.104	1.000				
Verified	-0.278	-0.072	-0.018	-0.003	-0.031	0.003	-0.061	-0.028	0.080	0.218	1.000			
AccountAge	-0.185	-0.035	-0.033	-0.020	-0.030	-0.007	-0.036	-0.019	-0.060	0.047	0.321	1.000		
Followers	-0.220	-0.045	0.000	-0.019	-0.015	0.028	-0.035	-0.014	0.104	0.116	0.429	0.227	1.000	
Followees	-0.006	-0.011	-0.001	0.091	-0.017	-0.002	-0.023	-0.001	-0.071	0.006	0.053	0.105	0.044	1.000

Table S2.7: VIFs

	News	Post
Falsehood	1.24	1.22
Conspiracy	1.05	1.04
Anger	1.07	1.05
Disgust	1.07	1.09
Fear	1.18	1.25
Joy	1.10	1.14
Sadness	1.13	1.21
Surprise	1.01	1.03
Words	1.19	1.09
Media		1.06
Verified		1.41
AccountAge		1.16
Followers		1.27
Followees		1.03
Mean VIF	1.12	1.15

S2.3 Robustness Check With Original 26 Topics

We adopt the original 26 topics and associated keywords directly from previous research [87] and repeat the analysis. Table S2.8 reports the estimation results for the topic diversity. Table S2.9 reports the estimation results for the engagement at news source sharer level. Table S2.10 reports the estimation results for the engagement at post viewer level. All the results are robust and consistently support our main findings.

	(1) News claim	(2) News post	(3) Claim → Post
Falsehood	0.404*** (0.034)	0.796*** (0.023)	0.724*** (0.022)
Conspiracy	0.030 (0.095)	0.117*** (0.015)	0.165*** (0.019)
Words	0.417*** (0.013)	0.393*** (0.003)	0.412*** (0.004)
Media		0.008 (0.006)	0.018* (0.008)
Verified		0.043*** (0.008)	0.013 (0.010)
AccountAge		0.002 (0.003)	−0.001 (0.003)
Followers		−0.001 (0.003)	−0.021*** (0.004)
Followees		−0.015*** (0.002)	−0.018*** (0.003)
Emotions	✓	✓	✓
Language	✓	✓	✓
MonthYear	✓	✓	✓
Intercept	−0.078 (0.109)	−0.709*** (0.037)	−0.873*** (0.046)
News-level REs	✗	✓	✓
<i>N</i>	7479	73,006	73,006
<i>R</i> ²	0.196	✗	✗

Table S2.8: Estimation results for the topic diversity in news claims [Column (1)], the topic diversity in news posts [Column (2)], and the changes in topic diversity from news claims to associated posts [Column (3)]. The topic diversity is recalculated based on the original 26 topics. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table S2.9: Estimation results for news lifetime [Column (1)] and post count [Column (2)]. The topic diversity is recalculated based on the original 26 topics. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) News lifetime	(2) Post count
Falsehood	1.232*** (0.045)	2.220*** (0.082)
Conspiracy	0.412* (0.186)	0.486 (0.332)
Diversity	−0.002 (0.005)	−0.045 (0.026)
Falsehood × Diversity	−0.069 (0.040)	0.084 (0.081)
Words	−0.081*** (0.010)	−0.609*** (0.030)
Emotions	✓	✓
Language	✓	✓
MonthYear	✓	✓
Intercept	2.534*** (0.274)	2.062*** (0.197)
<i>N</i>	7479	7479
<i>R</i> ²	0.478	0.185

Table S2.10: Estimation results for repost count [Column (1)], like count [Column (2)], and reply count [Column (3)]. The topic diversity is recalculated based on the original 26 topics. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) Repost count	(2) Like count	(3) Reply count
Falsehood	−0.240*** (0.061)	−0.150* (0.061)	0.407*** (0.057)
Conspiracy	0.335*** (0.064)	0.365*** (0.059)	0.361*** (0.056)
Diversity	0.160*** (0.030)	0.167*** (0.029)	0.118*** (0.029)
Falsehood × Diversity	−0.048 (0.034)	−0.034 (0.032)	−0.042 (0.033)
Words	0.517*** (0.017)	0.467*** (0.016)	0.442*** (0.015)
Media	0.774*** (0.027)	0.594*** (0.026)	0.426*** (0.026)
Verified	3.027*** (0.035)	3.036*** (0.034)	2.368*** (0.032)
AccountAge	0.094*** (0.012)	0.089*** (0.011)	0.044*** (0.011)
Followers	0.186*** (0.015)	0.184*** (0.014)	0.221*** (0.013)
Followees	0.464*** (0.022)	0.300*** (0.019)	0.189*** (0.014)
Emotions	✓	✓	✓
Language	✓	✓	✓
MonthYear	✓	✓	✓
Intercept	0.057 (0.165)	1.020*** (0.151)	−1.447*** (0.150)
News-level REs	✓	✓	✓
<i>N</i>	73,006	73,006	73,006
<i>R</i> ²	✗	✗	✗

S2.4 Robustness Check at Theme Level

We recalculate topic diversity at the theme level and repeat our analysis. Table S2.11 reports the estimation results for the topic diversity. Table S2.12 reports the estimation results for the engagement at news source sharer level. Table S2.13 reports the estimation results for the engagement at post viewer level. All the results are robust and consistently support our main findings.

	(1) News claim	(2) News post	(3) Claim → Post
Falsehood	0.311 *** (0.036)	0.699 *** (0.026)	0.616 *** (0.023)
Conspiracy	0.033 (0.092)	0.130 *** (0.016)	0.173 *** (0.019)
Words	0.354 *** (0.013)	0.337 *** (0.004)	0.342 *** (0.004)
Media		0.030 *** (0.007)	0.040 *** (0.008)
Verified		0.033 *** (0.008)	−0.002 (0.010)
AccountAge		0.005* (0.003)	0.004 (0.003)
Followers		0.002 (0.003)	−0.020 *** (0.004)
Followees		−0.014 *** (0.002)	−0.015 *** (0.003)
Emotions	✓	✓	✓
Language	✓	✓	✓
MonthYear	✓	✓	✓
Intercept	−0.101 (0.110)	−0.517 *** (0.040)	−0.600 *** (0.047)
News-level REs	✗	✓	✓
N	7273	70,904	70,904
R ²	0.147	✗	✗

Table S2.11: Estimation results for the topic diversity in news claims [Column (1)], the topic diversity in news posts [Column (2)], and the changes in topic diversity from news claims to associated posts [Column (3)]. The topic diversity is recalculated at the theme level. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table S2.12: Estimation results for news lifetime [Column (1)] and post count [Column (2)]. The topic diversity is recalculated at the theme level. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) News lifetime	(2) Post count
Falsehood	1.235*** (0.046)	2.238*** (0.084)
Conspiracy	0.448* (0.188)	0.485 (0.324)
Diversity	−0.005 (0.005)	−0.048 (0.026)
Falsehood × Diversity	−0.043 (0.042)	0.044 (0.081)
Words	−0.082*** (0.010)	−0.602*** (0.030)
Emotions	✓	✓
Language	✓	✓
MonthYear	✓	✓
Intercept	2.490*** (0.282)	2.076*** (0.199)
<i>N</i>	7273	7273
<i>R</i> ²	0.480	0.188

Table S2.13: Estimation results for repost count [Column (1)], like count [Column (2)], and reply count [Column (3)]. The topic diversity is recalculated at the theme level. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) Repost count	(2) Like count	(3) Reply count
Falsehood	−0.225*** (0.062)	−0.125* (0.062)	0.415*** (0.057)
Conspiracy	0.346*** (0.065)	0.378*** (0.059)	0.368*** (0.056)
Diversity	0.159*** (0.030)	0.175*** (0.028)	0.125*** (0.028)
Falsehood × Diversity	−0.124*** (0.034)	−0.094** (0.032)	−0.074* (0.031)
Words	0.528*** (0.017)	0.477*** (0.016)	0.444*** (0.015)
Media	0.786*** (0.028)	0.596*** (0.027)	0.431*** (0.027)
Verified	3.036*** (0.035)	3.070*** (0.034)	2.366*** (0.033)
AccountAge	0.091*** (0.012)	0.084*** (0.011)	0.041*** (0.012)
Followers	0.175*** (0.015)	0.180*** (0.014)	0.217*** (0.013)
Followees	0.546*** (0.024)	0.383*** (0.022)	0.207*** (0.015)
Emotions	✓	✓	✓
Language	✓	✓	✓
MonthYear	✓	✓	✓
Intercept	0.098 (0.169)	1.023*** (0.154)	−1.418*** (0.153)
News-level REs	✓	✓	✓
<i>N</i>	70,904	70,904	70,904
<i>R</i> ²	✗	✗	✗

S2.5 Robustness Check With Propensity Score Matching

To reduce potential unobserved confounding factors and ensure the robustness of our results, we further conduct propensity score matching for true and false news claims. Specifically, we match true and false news claims based on the variables of emotions and the number of words, with settings including the caliper of 0.3, “common support”, and “noreplacement.” Consequently, we discard 30 false news claims and achieve all biases lower than 0.05. Subsequently, we repeat our analysis based on the matched true and false news claims and their associated posts on $\mathbb{X}$. Table S2.14 reports the estimation results for the topic diversity. Table S2.15 reports the estimation results for the engagement at news source sharer level. Table S2.16 reports the estimation results for the engagement at post viewer level. All the results are robust and consistently support our main findings.

	(1) News claim	(2) News post	(3) Claim → Post
Falsehood	0.400*** (0.036)	0.745*** (0.024)	0.640*** (0.022)
Conspiracy	0.037 (0.091)	0.137*** (0.015)	0.186*** (0.019)
Words	0.387*** (0.013)	0.369*** (0.004)	0.383*** (0.004)
Media		0.021*** (0.006)	0.034*** (0.008)
Verified		0.041*** (0.008)	0.009 (0.010)
AccountAge		0.003 (0.003)	0.001 (0.003)
Followers		−0.001 (0.003)	−0.019*** (0.004)
Followees		−0.010*** (0.002)	−0.011*** (0.003)
Emotions	✓	✓	✓
Language	✓	✓	✓
MonthYear	✓	✓	✓
Intercept	−0.075 (0.111)	−0.644*** (0.039)	−0.705*** (0.047)
News-level REs	✗	✓	✓
<i>N</i>	7243	69,366	69,366
<i>R</i> ²	0.181	✗	✗

Table S2.14: Estimation results for the topic diversity in news claims [Column (1)], the topic diversity in news posts [Column (2)], and the changes in topic diversity from news claims to associated posts [Column (3)]. The estimations are based on samples obtained after propensity score matching. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table S2.15: Estimation results for news lifetime [Column (1)] and post count [Column (2)]. The estimations are based on samples obtained after propensity score matching. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) News lifetime	(2) Post count
Falsehood	1.234*** (0.046)	2.230*** (0.083)
Conspiracy	0.479* (0.189)	0.501 (0.337)
Diversity	0.000 (0.005)	−0.024 (0.026)
Falsehood × Diversity	−0.073 (0.040)	0.077 (0.086)
Words	−0.082*** (0.010)	−0.615*** (0.030)
Emotions	✓	✓
Language	✓	✓
MonthYear	✓	✓
Intercept	2.572*** (0.275)	2.077*** (0.197)
<i>N</i>	7243	7243
<i>R</i> ²	0.482	0.188

Table S2.16: Estimation results for repost count [Column (1)], like count [Column (2)], and reply count [Column (3)]. The estimations are based on samples obtained after propensity score matching. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) Repost count	(2) Like count	(3) Reply count
Falsehood	−0.244*** (0.063)	−0.134* (0.063)	0.411*** (0.058)
Conspiracy	0.360*** (0.065)	0.391*** (0.059)	0.370*** (0.057)
Diversity	0.183*** (0.031)	0.193*** (0.030)	0.133*** (0.030)
Falsehood × Diversity	−0.125*** (0.035)	−0.088** (0.033)	−0.088** (0.033)
Words	0.512*** (0.017)	0.459*** (0.016)	0.442*** (0.015)
Media	0.788*** (0.028)	0.597*** (0.027)	0.432*** (0.027)
Verified	3.075*** (0.036)	3.118*** (0.035)	2.405*** (0.033)
AccountAge	0.088*** (0.012)	0.081*** (0.011)	0.038** (0.012)
Followers	0.172*** (0.015)	0.177*** (0.014)	0.216*** (0.013)
Followees	0.635*** (0.026)	0.467*** (0.024)	0.250*** (0.017)
Emotions	✓	✓	✓
Language	✓	✓	✓
MonthYear	✓	✓	✓
Intercept	0.061 (0.169)	0.974*** (0.155)	−1.475*** (0.154)
News-level REs	✓	✓	✓
<i>N</i>	69,366	69,366	69,366
<i>R</i> ²	✗	✗	✗

S2.6 Robustness Check With Inclusion of QAnon Conspiracy Theory

We further include QAnon to extend conspiracy theories and repeat our analysis. QAnon is a far-right conspiracy theory alleging that a cabal of politicians and media personalities are engaged in a war to undermine society [94, 129, 130]. The QAnon discourse has become more diverse and evolved into a global phenomenon during the COVID-19 pandemic [132]. We adopt a list of QAnon-specific keywords from previous research [133] to identify QAnon conspiracy theory in our dataset. The performance of this keyword list in identifying QAnon discourse has been manually validated, showing high agreement with human assessments [130]. As a result, we identify 6 QAnon news claims and 302 QAnon news posts. Table S2.17 reports the estimation results for the topic diversity. Table S2.18 reports the estimation results for the engagement at news source sharer level. Table S2.19 reports the estimation results for the engagement at post viewer level. All the results are robust and consistently support our main findings.

	(1) News claim	(2) News post	(3) Claim → Post
Falsehood	0.389*** (0.036)	0.751*** (0.024)	0.659*** (0.022)
Conspiracy	0.030 (0.087)	0.124*** (0.015)	0.175*** (0.018)
Words	0.391*** (0.013)	0.373*** (0.004)	0.389*** (0.004)
Media		0.022*** (0.006)	0.034*** (0.008)
Verified		0.037*** (0.008)	0.005 (0.010)
AccountAge		0.003 (0.003)	0.001 (0.003)
Followers		−0.001 (0.003)	−0.019*** (0.004)
Followees		−0.013*** (0.002)	−0.015*** (0.003)
Emotions	✓	✓	✓
Language	✓	✓	✓
MonthYear	✓	✓	✓
Intercept	−0.075 (0.110)	−0.652*** (0.039)	−0.717*** (0.047)
News-level REs	✗	✓	✓
<i>N</i>	7273	70,904	70,904
<i>R</i> ²	0.179	✗	✗

Table S2.17: Estimation results for the topic diversity in news claims [Column (1)], the topic diversity in news posts [Column (2)], and the changes in topic diversity from news claims to associated posts [Column (3)]. QAnon conspiracy theory is included during estimation. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table S2.18: Estimation results for news lifetime [Column (1)] and post count [Column (2)]. QAnon conspiracy theory is included during estimation. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) News lifetime	(2) Post count
Falsehood	1.234*** (0.046)	2.232*** (0.083)
Conspiracy	0.505** (0.183)	0.443 (0.330)
Diversity	-0.002 (0.005)	-0.025 (0.026)
Falsehood × Diversity	-0.074 (0.039)	0.080 (0.085)
Words	-0.080*** (0.010)	-0.613*** (0.030)
Emotions	✓	✓
Language	✓	✓
MonthYear	✓	✓
Intercept	2.487*** (0.282)	2.065*** (0.198)
<i>N</i>	7273	7273
<i>R</i> ²	0.481	0.188

Table S2.19: Estimation results for re-post count [Column (1)], like count [Column (2)], and reply count [Column (3)]. QAnon conspiracy theory is included during estimation. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) Repost count	(2) Like count	(3) Reply count
Falsehood	-0.251*** (0.063)	-0.150* (0.063)	0.403*** (0.058)
Conspiracy	0.271*** (0.063)	0.300*** (0.058)	0.312*** (0.055)
Diversity	0.181*** (0.031)	0.192*** (0.030)	0.137*** (0.030)
Falsehood × Diversity	-0.125*** (0.035)	-0.092** (0.033)	-0.090** (0.033)
Words	0.520*** (0.017)	0.468*** (0.016)	0.443*** (0.015)
Media	0.788*** (0.028)	0.596*** (0.027)	0.432*** (0.027)
Verified	3.037*** (0.035)	3.071*** (0.034)	2.367*** (0.033)
AccountAge	0.092*** (0.012)	0.085*** (0.011)	0.041*** (0.012)
Followers	0.175*** (0.015)	0.179*** (0.014)	0.217*** (0.013)
Followees	0.545*** (0.024)	0.383*** (0.022)	0.207*** (0.015)
Emotions	✓	✓	✓
Language	✓	✓	✓
MonthYear	✓	✓	✓
Intercept	0.130 (0.169)	1.055*** (0.155)	-1.396*** (0.153)
News-level REs	✓	✓	✓
<i>N</i>	70,904	70,904	70,904
<i>R</i> ²	✗	✗	✗

S2.7 Robustness Check With Exclusion of Potential Debunking Posts

News posts can also debunk the linked news items. We use keywords of “misinformation,” “misleading,” “fake,” “false,” and “disinformation” to identify potential debunking posts, resulting in 4316 news posts. We exclude these potential debunking posts and repeat our analysis. Table S2.20 reports the estimation results for the topic diversity. Table S2.21 reports the estimation results for the engagement at news source sharer level. Table S2.22 reports the estimation results for the engagement at post viewer level. All the results are robust and consistently support our main findings.

	(1) News claim	(2) News post	(3) Claim → Post
Falsehood	0.364*** (0.037)	0.773*** (0.025)	0.717*** (0.023)
Conspiracy	0.065 (0.095)	0.124*** (0.016)	0.171*** (0.020)
Words	0.381*** (0.013)	0.370*** (0.004)	0.384*** (0.004)
Media		0.021** (0.007)	0.033*** (0.008)
Verified		0.049*** (0.009)	0.018 (0.011)
AccountAge		0.002 (0.003)	0.001 (0.003)
Followers		−0.005 (0.003)	−0.026*** (0.004)
Followees		−0.009*** (0.002)	−0.010** (0.003)
Emotions	✓	✓	✓
Language	✓	✓	✓
MonthYear	✓	✓	✓
Intercept	−0.104 (0.117)	−0.636*** (0.039)	−0.697*** (0.048)
News-level REs	✗	✓	✓
N	7097	66,588	66,588
R ²	0.173	✗	✗

Table S2.20: Estimation results for the topic diversity in news claims [Column (1)], the topic diversity in news posts [Column (2)], and the changes in topic diversity from news claims to associated posts [Column (3)]. The potential debunking posts are excluded during estimation. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table S2.21: Estimation results for news lifetime [Column (1)] and post count [Column (2)]. The potential debunking posts are excluded during estimation. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) News lifetime	(2) Post count
Falsehood	1.307*** (0.049)	2.234*** (0.087)
Conspiracy	0.401* (0.203)	0.463 (0.344)
Diversity	−0.004 (0.005)	−0.023 (0.026)
Falsehood × Diversity	−0.065 (0.044)	0.077 (0.093)
Words	−0.078*** (0.010)	−0.619*** (0.029)
Emotions	✓	✓
Language	✓	✓
MonthYear	✓	✓
Intercept	2.599*** (0.288)	2.117 *** (0.198)
<i>N</i>	7097	7097
<i>R</i> ²	0.496	0.191

Table S2.22: Estimation results for repost count [Column (1)], like count [Column (2)], and reply count [Column (3)]. The potential debunking posts are excluded during estimation. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) Repost count	(2) Like count	(3) Reply count
Falsehood	−0.207** (0.065)	−0.050 (0.065)	0.525*** (0.060)
Conspiracy	0.295*** (0.069)	0.320*** (0.063)	0.365*** (0.059)
Diversity	0.177*** (0.032)	0.191*** (0.030)	0.138*** (0.030)
Falsehood × Diversity	−0.118 *** (0.036)	−0.086* (0.034)	−0.101 ** (0.033)
Words	0.527*** (0.017)	0.471*** (0.016)	0.454*** (0.015)
Media	0.768*** (0.029)	0.581*** (0.028)	0.416 *** (0.027)
Verified	3.071*** (0.038)	3.127*** (0.037)	2.420*** (0.035)
AccountAge	0.097*** (0.013)	0.083*** (0.012)	0.036** (0.012)
Followers	0.171 *** (0.015)	0.174 *** (0.014)	0.208*** (0.013)
Followees	0.604*** (0.025)	0.440*** (0.023)	0.238*** (0.016)
Emotions	✓	✓	✓
Language	✓	✓	✓
MonthYear	✓	✓	✓
Intercept	0.160 (0.174)	1.050*** (0.159)	−1.452*** (0.159)
News-level REs	✓	✓	✓
<i>N</i>	66,588	66,588	66,588
<i>R</i> ²	X	X	X

S2.8 Robustness Check With BERTopic

To ensure the generalization of our results across the topic modeling methods, we use BERTopic, a popular unsupervised topic modeling tool, as an alternative method to automatically cluster topics.¹ The BERTopic model generates 1526 topic clusters for the whole news claims and posts. We use the topic clusters and their corresponding probabilities to calculate topic diversity again and repeat the analysis. Table S2.23 reports the estimation results for the topic diversity. Table S2.24 reports the estimation results for the engagement at news source sharer level. Table S2.25 reports the estimation results for the engagement at post viewer level. All the results are robust and consistently support our main findings.

1: <https://maartengr.github.io/BERTopic/index.html>

	(1) News claim	(2) News post	(3) Claim → Post
Falsehood	−0.038 (0.042)	−0.042 (0.028)	0.059** (0.021)
Conspiracy	−0.150 (0.127)	0.046* (0.020)	0.045* (0.018)
Words	0.089*** (0.013)	0.144*** (0.005)	0.110*** (0.004)
Media		0.016 (0.009)	0.017* (0.007)
Verified		0.016 (0.011)	−0.016 (0.009)
AccountAge		0.001 (0.003)	0.000 (0.003)
Followers		−0.009* (0.004)	−0.007* (0.004)
Followees		−0.019*** (0.003)	−0.016*** (0.003)
Emotions	✓	✓	✓
Language	✓	✓	✓
MonthYear	✓	✓	✓
Intercept	−0.162 (0.160)	0.222*** (0.051)	−0.042 (0.044)
News-level REs	✗	✓	✓
<i>N</i>	7273	70,904	70,904
<i>R</i> ²	0.041	✗	✗

Table S2.23: Estimation results for the topic diversity in news claims [Column (1)], the topic diversity in news posts [Column (2)], and the changes in topic diversity from news claims to associated posts [Column (3)]. The topic diversity is calculated based on topics generated by BERTopic. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table S2.24: Estimation results for news lifetime [Column (1)] and post count [Column (2)]. The topic diversity is calculated based on topics generated by BERTopic. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) News lifetime	(2) Post count
Falsehood	1.209*** (0.044)	2.175*** (0.084)
Conspiracy	0.405* (0.186)	0.553 (0.350)
Diversity	-0.026*** (0.005)	-0.178*** (0.027)
Falsehood $\times$ Diversity	-0.302*** (0.034)	-0.117 (0.068)
Words	-0.086*** (0.010)	-0.596*** (0.029)
Emotions	✓	✓
Language	✓	✓
MonthYear	✓	✓
Intercept	2.394*** (0.275)	2.029*** (0.202)
N	7273	7273
R^2	0.508	0.192

Table S2.25: Estimation results for repost count [Column (1)], like count [Column (2)], and reply count [Column (3)]. The topic diversity is calculated based on topics generated by BERTopic. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) Repost count	(2) Like count	(3) Reply count
Falsehood	-0.134* (0.060)	0.000 (0.060)	0.521*** (0.056)
Conspiracy	0.351*** (0.065)	0.387*** (0.059)	0.375*** (0.056)
Diversity	0.132*** (0.024)	0.197*** (0.023)	0.211*** (0.024)
Falsehood $\times$ Diversity	-0.058* (0.028)	-0.094*** (0.026)	-0.173*** (0.027)
Words	0.537*** (0.016)	0.492*** (0.015)	0.450*** (0.014)
Media	0.783*** (0.028)	0.591*** (0.027)	0.432*** (0.027)
Verified	3.041*** (0.035)	3.076*** (0.034)	2.369*** (0.033)
AccountAge	0.092*** (0.012)	0.086*** (0.011)	0.043*** (0.012)
Followers	0.174*** (0.015)	0.181*** (0.014)	0.218*** (0.013)
Followees	0.550*** (0.024)	0.389*** (0.022)	0.207*** (0.015)
Emotions	✓	✓	✓
Language	✓	✓	✓
MonthYear	✓	✓	✓
Intercept	0.013 (0.168)	0.919*** (0.154)	-1.506*** (0.152)
News-level REs	✓	✓	✓
N	70,904	70,904	70,904
R^2	$\times$	$\times$	$\times$

S2.9 Additional Analysis on Quote Count

Table S2.26 reports the full estimation results for quote count.

	(1) Quote count
Falsehood	0.102 (0.074)
Conspiracy	0.585*** (0.084)
Diversity	0.186*** (0.039)
Falsehood × Diversity	−0.150*** (0.045)
Words	0.414*** (0.021)
Media	0.958*** (0.037)
Verified	3.106*** (0.044)
AccountAge	0.104*** (0.018)
Followers	0.160*** (0.014)
Followees	0.294*** (0.022)
Emotions	✓
Language	✓
MonthYear	✓
Intercept	−1.981*** (0.218)
News-level REs	✓
<i>N</i>	70,904
<i>R</i> ²	X

Table S2.26: Estimation results for quote count. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

S3

Supplementary Materials for Chapter 3

S3.1 Dataset Overview

An overview of the dataset used in this study is shown in Table S3.1.

	(1) All	(2) PolitiFact	(3) Snopes
Number of statements	35,014	22,561	12,453
Observation start date	01/02/2008	01/02/2008	01/04/2008
Observation end date	07/19/2023	07/19/2023	07/17/2023
Number of fact-checkers	608	585	23
Number of topics	202	152	66
Politician	9,308	6,363	2,945
	(26.6%)	(28.2%)	(23.6%)
Republican	4,403	2,881	1,522
	(12.6%)	(12.8%)	(12.2%)
Democrat	4,455	3,207	1,248
	(12.7%)	(14.2%)	(10.0%)
Veracity			
True	9,616	6,199	3,417
	(27.5%)	(27.5%)	(27.4%)
False	25,398	16,362	9,036
	(72.5%)	(72.5%)	(72.6%)
Months to elections	11.603	11.078	12.555
	(0.039)	(0.048)	(0.067)
Other-condemning emotions	0.096	0.092	0.103
	(0.001)	(0.001)	(0.002)
Words	18.376	18.237	18.628
	(0.042)	(0.054)	(0.064)

Table S3.1: Dataset overview. Column (1) reports summary statistics for the whole dataset. Column (2) reports summary statistics for the subset of fact-checked statements from PolitiFact. Column (3) reports summary statistics for the subset of fact-checked statements from Snopes. Percentages or standard errors are included in the parentheses.

S3.2 Manual Validation

In addition to manually checking 200 random samples, we further use three steps to check the precision of our politician matching approach for all politicized statements:

Step 1: We additionally extract the fact-checking comments in the corresponding fact-checking webpages. Fact-checkers provide these comments to explain their verdicts for the veracity of the statements. If the comments also mention other politicians except the politicians mentioned in the original statements. Then the politician names matched in the statements have a high probability of being corresponding politicians. Given this, we match politician names mentioned in the comments. If there is at least one other politician in the comments, we take the associated statements as true positives. As a result, out of the total fact-checked statements that mention politicians, 88.5% are linked to the comments that mention other politicians.

Step 2: We create a list of specific politicians, including Joe Biden, Kamala Harris, Donald Trump, Barack Obama, Nancy Pelosi, and Hillary Clinton.

Table S3.2: The twenty most frequent topic tags attached to fact-checked statements. Column (1) reports the number of fact-checked true statements for each topic. Column (2) reports the number of fact-checked false statements for each topic. Column (3) reports the number of all fact-checked statements for each topic.

	(1) True	(2) False	(3) Total
Facebook fact-checks	326	4,009	4,335
Politics	1,315	2,663	3,978
Fauxtography	330	1,669	1,999
Health care	528	1,453	1,981
Economy	699	1,130	1,829
Elections	448	1,258	1,706
Taxes	517	1,107	1,624
Coronavirus	156	1,334	1,490
Junk news	3	1,412	1,415
Crime	381	957	1,338
Education	461	845	1,306
Immigration	287	935	1,222
Public health	226	953	1,179
Jobs	391	750	1,141
Federal budget	370	743	1,113
State budget	388	645	1,033
Candidate biography	297	699	996
History	362	607	969
Fake news	17	839	856
Foreign policy	231	584	815

We take statements that mention these specific names as true positives, because these politician names are somewhat specific to the corresponding politicians in the context of fact-checking. The true positives in this regard are 768 out of 1070 statements that are not identified by *Step 1*.

Step 3: We manually check the remaining 302 statements from Steps 1 and 2. We find there are 33 false positives that mention the names same as politician names but do not correspond to the politicians. Finally, we combine the three steps together and find that our politician matching approach has high performance with a precision score of 99.6%.

S3.3 Analysis Across Topics

The fact-checked statements in our data concern a wide spectrum of 202 topics (see Section 3.4). Notably, out of the 202 topics, 194 (96%) are present in both true and false statements. Table S3.2 displays the 20 most frequent topics shared by fact-checked true and false statements. Furthermore, as shown in the embedded figure within Fig. S3.1a, the ratios of topics in fact-checked true statements have significantly positive correlations with the ratios of corresponding topics in fact-checked false statements ($r = 0.751, p < 0.001$). Additionally, using a paired two-sample t -test ($t = 0.009, p > 0.05$), we find that the ratios of each shared topic in fact-checked true and false statements do not differ significantly (Fig. S3.1a). This suggests that the distribution of statements across topics is not veracity-specific.

While the distribution of topics appears to be independent of veracity (Fig. S3.1a), the degree of politicization within fact-checked true and false claims may differ across social issues. We examine how mentions of political elites in fact-checked true and false statements vary across topics. Given that a single statement can discuss multiple topics, we assess each topic within the same statement separately. To this end, we count all

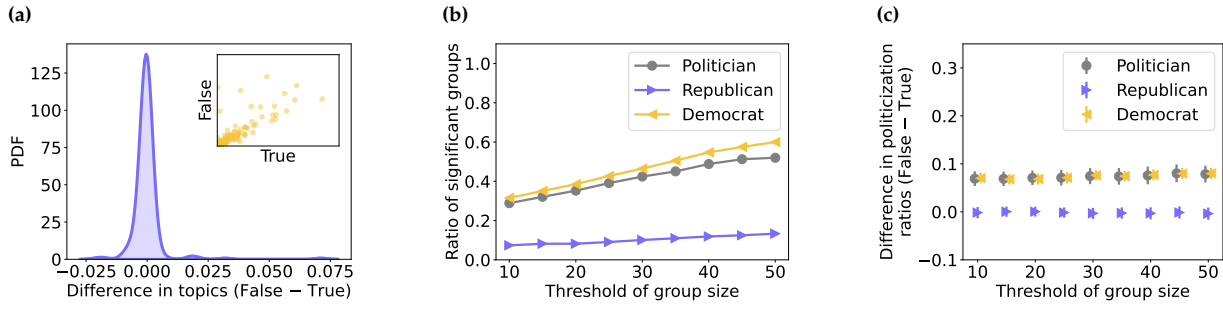


Figure S3.1: Topics in fact-checked statements. (a) The distribution of the differences in shared topics between fact-checked true and false statements. The embedded scatters describe the ratios of each shared topic in fact-checked true and false statements. (b) The ratios of statement groups in which the mentions of political elites are significantly associated with the veracity of fact-checked statements under different thresholds of group size. (c) The differences in politicization ratios between fact-checked true and false statements under different thresholds of group size. The error bars represent 95% Confidence Intervals (CIs).

fact-checked statements that share the same topic as one statement group. Within each topic-specific statement group, we employ a Chi-Square test to evaluate the association between politicization and veracity. For these in-group tests, it's worth noting that some groups may lack a sufficient number of samples. Therefore, we set a group size threshold ranging from 10 to 50 (in increments of 5) to assess within-group differences. In each Chi-Square test, the input is a 2×2 matrix including the counts of politicized and non-politicized statements in true and false categories.

As shown in Fig. S3.1b, the ratio of statement groups in which mentions of political elites exhibit a statistically significant relationship with veracity relative to all statement groups increases from 28.9% to 52.0% with increasing group size thresholds. However, notably, the ratio of statement groups in which mentions of Republicans show a significant association with veracity relative to all statement groups remains relatively low, ranging from 7.4% to 13.3%. Conversely, the ratio of statement groups in which mentions of Democrats show a significant relationship with veracity relative to all statement groups rises from 31.5% to 60.0%. These findings suggest that across various topics, fact-checked false and true statements consistently show opposing tendencies when it comes to mentioning political elites, particularly Democrats.

We further consider the politicization ratio within each group and conduct out-group paired *t*-tests between fact-checked true and false statements under each group size threshold. The politicization ratio indicates the ratio of statements mentioning political elites relative to all statements in each group. Fig. S3.1c shows the variance of the difference in politicization ratio between true and false statements among the topic-specific groups under each threshold for all politicians (grey), Republicans (purple), and Democrats (yellow). We consistently observe that the politicization ratio in fact-checked false statements is significantly higher than that in fact-checked true statements. This disparity primarily arises from the increased likelihood of fact-checked false statements mentioning Democrats, while no statistically significant difference between fact-checked true and false statements is found with regard to mentions of Republicans (Fig. S3.1c).

Additionally, we keep the group size threshold at 50 and examine the changes in politicization ratios of individual topics between Obama's incumbency and Trump's incumbency. Specifically, 59 topics are presented in both Obama's incumbency and Trump's incumbency, and each has at least 50 fact-checked statements during either Obama's incumbency

or Trump's incumbency. There are only nine topics with significantly different politicization ratios between Obama's incumbency and Trump's incumbency (as shown in Fig. 3.3). The other shared topics have similar politicization ratios during two incumbencies.

S3.4 Analysis Across Incumbent U.S. Presidents

In the main analysis, we incorporate incumbent fixed effect (*Trump*) and its interactions with *Falsehood* and *Politician* to examine the differences in mentions of politicians and other-condemning emotions between Obama's incumbency and Trump's incumbency. The estimation results for the main analysis of mentions of political elites and other-condemning emotions across incumbencies are reported in Table S3.3 and Column (1) of Table S3.5, respectively. Subsequently, we further conduct subgroup analysis across incumbent presidents. The estimation results for the mentions of political elites are reported in Table S3.4. Columns (1)–(3) of Table S3.4 report the estimation results for fact-checked statements during Obama's incumbency, and Columns (4)–(6) report the results for fact-checked statements during Trump's incumbency. These results are consistent with our findings in Table S3.3. Additionally, in Table S3.5, the estimation results for the other-condemning emotions in the fact-checked statement subgroups during Obama's [Columns (2)–(4)] and Trump's incumbencies [Columns (5)–(7)] are consistent with the interaction analysis in Column (1).

Table S3.3: Estimation results for mixed logistic regression models predicting whether the fact-checked statements mention political elites of either political party [Column (1)], Republicans [Column (2)], or Democrats [Column (3)]. Fact-checker-specific random effects and incumbency (Trump) fixed effect are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) Politician	(2) Republican	(3) Democrat
Falsehood	0.169** (0.064)	−0.297*** (0.077)	0.634*** (0.091)
MTE	−0.073*** (0.004)	−0.063*** (0.005)	−0.060*** (0.007)
Falsehood × MTE	0.028*** (0.005)	0.039*** (0.006)	0.010 (0.007)
Words	0.016*** (0.002)	0.012*** (0.002)	0.013*** (0.002)
Trump	0.174* (0.069)	0.133 (0.082)	0.105 (0.101)
Falsehood × Trump	−0.134 (0.072)	−0.258** (0.087)	0.039 (0.106)
Intercept	−1.055*** (0.080)	−1.667*** (0.093)	−2.209*** (0.100)
Fact-checker REs	✓	✓	✓
#Statements	25,967	25,967	25,967

Table S3.4: Estimation results for mixed logistic regression models predicting whether the fact-checked statements mention political elites of either political party [Columns (1) and (4)], Republicans [Columns (2) and (5)], or Democrats [Columns (3) and (6)]. Columns (1)–(3) report results based on the fact-checked statements during Obama’s incumbency. Columns (4)–(6) report results based on the fact-checked statements during Trump’s incumbency. Fact-checker-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	Obama			Trump		
	(1) Politician	(2) Republican	(3) Democrat	(4) Politician	(5) Republican	(6) Democrat
Falsehood	0.326*** (0.075)	−0.122 (0.090)	0.655*** (0.104)	−0.061 (0.098)	−0.689*** (0.118)	0.684*** (0.143)
MTE	−0.088*** (0.006)	−0.075*** (0.007)	−0.072*** (0.009)	−0.051*** (0.007)	−0.045*** (0.008)	−0.040*** (0.011)
Falsehood × MTE	0.012 (0.007)	0.021* (0.008)	0.007 (0.010)	0.034*** (0.008)	0.049*** (0.009)	0.007 (0.012)
Words	0.007** (0.003)	−0.004 (0.003)	0.015*** (0.003)	0.026*** (0.003)	0.028*** (0.003)	0.011*** (0.003)
Intercept	−0.697*** (0.099)	−1.252*** (0.112)	−2.136*** (0.120)	−1.311*** (0.124)	−2.011*** (0.142)	−2.290*** (0.159)
Fact-checker REs	✓	✓	✓	✓	✓	✓
#Statements	14,515	14,515	14,515	11,452	11,452	11,452

Table S3.5: Estimation results for mixed linear regression models predicting other-condemning emotions in the fact-checked statements. Column (1) reports result for the model that incorporates incumbency (Trump) fixed effect. Columns (2)–(4) report results based on the fact-checked statements during Obama’s incumbency. Columns (5)–(7) report results based on the fact-checked statements during Trump’s incumbency. Fact-checker-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	Obama				Trump		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Politician	0.012*** (0.003)	0.011*** (0.003)	0.019* (0.008)	0.037* (0.015)	0.004 (0.004)	0.009 (0.011)	0.006 (0.018)
Republican				−0.017 (0.014)			−0.007 (0.016)
Democrat				−0.021 (0.014)			0.018 (0.016)
Falsehood	0.007* (0.003)	0.006 (0.003)	0.005 (0.004)	0.005 (0.004)	0.012** (0.004)	0.014** (0.005)	0.014** (0.005)
Politician × Falsehood			0.004 (0.007)	0.004 (0.007)		−0.007 (0.010)	−0.012 (0.010)
MTE	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.001* (0.000)	0.001 (0.000)	0.001 (0.000)
Politician × MTE			−0.001** (0.000)	−0.001** (0.000)		0.000 (0.001)	0.000 (0.001)
Words	0.001*** (0.000)	0.001*** (0.000)	0.001*** (0.000)	0.001*** (0.000)	0.001*** (0.000)	0.001*** (0.000)	0.001*** (0.000)
Trump	0.014** (0.005)						
Politician × Trump	−0.008 (0.005)						
Falsehood × Trump	0.005 (0.005)						
Intercept	0.059*** (0.004)	0.060*** (0.005)	0.056*** (0.006)	0.056*** (0.006)	0.071*** (0.007)	0.070*** (0.007)	0.070*** (0.007)
Fact-checker REs	✓	✓	✓	✓	✓	✓	✓
#Statements	25,967	14,515	14,515	14,515	11,452	11,452	11,452

S3.5 Political Leanings

We first examine the partisan leanings based on all fact-checked statements. The estimation results are reported in Column (1) of Table S3.6 and used for the main analysis. Additionally, we conduct the same analysis again based on fact-checked statements from PolitiFact. The coefficient estimate of *Falsehood* in Column (2) of Table S3.6 is also significantly positive (*coef.* = 0.008, $p < 0.001$). Within fact-checked statements from PolitiFact, the predicted marginal means for partisan leanings in true and false statements are -0.004 (95% CI: $[-0.007, -0.001]$) and 0.004 (95% CI: $[0.001, 0.006]$), respectively. The estimates for partisan leanings of fact-checked true and false statements in PolitiFact are consistent with the estimates based on all fact-checked statements.

Table S3.6: Estimation results for mixed linear regression models predicting partisan leanings in the fact-checked statements. Column (1) reports result based on all fact-checked statements. Columns (2)–(4) report results based on fact-checked statements from PolitiFact. Fact-checker-specific random effects and yearly fixed effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) All	(2) PolitiFact	(3) PolitiFact	(4) PolitiFact
Falsehood	0.009*** (0.001)	0.008*** (0.002)	0.003* (0.002)	0.004 (0.003)
MTE	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Democrat _{author}			−0.021*** (0.002)	−0.032*** (0.005)
Republican _{author}			0.012*** (0.002)	0.013** (0.005)
Democrat _{author} × Falsehood				−0.014** (0.004)
Republican _{author} × Falsehood				0.011* (0.004)
Democrat _{author} × MTE				0.002*** (0.000)
Republican _{author} × MTE				−0.001*** (0.000)
Words	0.000 (0.000)	0.000* (0.000)	0.000* (0.000)	0.000* (0.000)
Intercept	−0.038*** (0.005)	−0.049*** (0.006)	−0.042*** (0.006)	−0.040*** (0.007)
Fact-checker REs	✓	✓	✓	✓
Year FEs	✓	✓	✓	
#Statements	35,014	22,561	22,561	22,561

Subsequently, given that PolitiFact provides author information for the fact-checked statements, we partially identify the political elites who authored the fact-checked statements based on our politician database and incorporate the author information into the regression. The estimation results are reported in Column (3) of Table S3.6 and used for the main analysis.

We further examine the moderating effects of *Falsehood* and *MTE* on the partisan leanings in fact-checked statements from Democrats and Republicans. As shown in Column (4) of Table S3.6, the coefficient estimate of *Democrat_{author} × Falsehood* is significantly negative (*coef.* = -0.014 , $p <$

0.01), and the coefficient estimate of $Republican_{author} \times Falsehood$ is significantly positive ($coef. = 0.011, p < 0.05$). This indicates that the in-group leanings for Republicans and Democrats are stronger in fact-checked false statements than in fact-checked true statements. Notably, the coefficient estimate of $Falsehood$ in Column (4) is not statistically significant ($coef. = 0.004, p > 0.05$). This suggests that the right-leaning of fact-checked false statements compared to fact-checked true statements is moderated by partisan authors.

The coefficient for $Democrat_{author} \times MTE$ is significantly positive ($coef. = 0.002, p < 0.001$), and the coefficient estimate of $Republican_{author} \times MTE$ is significantly negative ($coef. = -0.001, p < 0.001$). This suggests the tendency that fact-checked statements from political lefts and rights harness more other-condemning emotions toward political out-groups when elections approach. Finally, we specifically analyze whether fact-checked false statements from Democrats express other-condemning emotions differently compared to fact-checked false statements from Republicans, when mentioning political out-groups. The predicted marginal mean for partisan leaning of fact-checked false statements from Republicans is 0.018 (95% CI: [0.015, 0.021]), while the predicted marginal mean for partisan leaning of fact-checked false statements from Democrats is -0.022 (95% CI: [-0.026, -0.017]). The magnitudes of the predicted marginal means in fact-checked false statements from Democrats and Republicans are similar, which indicates that the pattern of harnessing animosity toward political out-groups is shared by both Republicans and Democrats in their fact-checked misinformation stories.

S3.6 Robustness Checks

S3.6.1 Cross-Correlations & Variance Inflation Factors

The cross-correlations of the independent variables and their variance inflation factors are shown in Table S3.7 and Table S3.8, respectively.

	Politician	Falsehood	MTE	Words
Politician	1.000			
Falsehood	0.050	1.000		
MTE	-0.129	-0.063	1.000	
Words	0.069	-0.076	0.019	1.000

Table S3.7: Cross-correlations among independent variables.

	VIF	1/VIF
Politician	1.02	0.976
Falsehood	1.01	0.988
MTE	1.02	0.980
Words	1.01	0.988

Table S3.8: Variance Inflation Factors (VIFs) for the independent variables.

S3.6.2 Statements Without Mixed Veracity

Statements with mixed veracity ratings (*e.g.*, half-true and half-false) are included and categorized as false in the main analysis. To ensure the robustness of our findings, we exclude these statements and perform our regression models again with the remaining 29,596 fact-checked true or

false statements. The regression results are reported in Table S3.9 and Table S3.10. We consistently find that fact-checked false statements are more likely to mention political elites, especially Democrats, compared to fact-checked true statements.

Table S3.9: Estimation results for mixed logistic regression models predicting whether the fact-checked statements mention political elites of either political party [Column (1)], Republicans [Column (2)], or Democrats [Column (3)]. Fact-checker-specific random effects and yearly fixed effects are included. The statements with mixed veracity are excluded. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) Politician	(2) Republican	(3) Democrat
Falsehood	0.146** (0.056)	−0.459*** (0.069)	0.690*** (0.078)
MTE	−0.041*** (0.004)	−0.033*** (0.005)	−0.042*** (0.006)
Falsehood × MTE	0.016*** (0.004)	0.026*** (0.005)	0.005 (0.006)
Words	0.024*** (0.002)	0.020*** (0.002)	0.017*** (0.002)
Intercept	−1.300*** (0.147)	−2.015*** (0.183)	−2.380*** (0.177)
Fact-checker REs	✓	✓	✓
Year FEs	✓	✓	✓
#Statements	29,596	29,596	29,596

S3.6.3 Statements With Manual Validation

We correct the false positives identified in Suppl. S3.2 and conduct the analysis again. The results remain robust (see Table S3.11 and Table S3.12 for details).

S3.6.4 Statements Without Mixed Political Elites

We exclude statements that mention two or more political elites with mixed party affiliations (*i.e.*, Democrats and Republicans) and conduct the analysis again. The results are robust and consistent with those in the main analysis (see Table S3.13 and Table S3.14 for details).

S3.6.5 Organization-Specific Fixed Effects

We repeat our analysis with organization-specific fixed effects. The results are reported in Table S3.15 and Table S3.16. Additionally, we repeat our analysis with fact-checker-specific fixed effects, which allows us to control the heterogeneity of fact-checkers explicitly. The results remain robust and are reported in Table S3.17 and Table S3.18.

S3.6.6 Author-Specific Fixed Effects

We repeat our analysis with author-specific fixed effects to control for the types of authors, *i.e.*, Republicans, Democrats, or others. The results remain robust and are reported in Table S3.19 and Table S3.20.

	(1)	(2)	(3)
Politician	0.008** (0.002)	0.019** (0.006)	0.023* (0.011)
Republican			−0.004 (0.010)
Democrat			−0.004 (0.010)
Falsehood	0.011*** (0.002)	0.013*** (0.003)	0.013*** (0.003)
Politician × Falsehood		−0.008 (0.005)	−0.008 (0.005)
MTE	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Politician × MTE		−0.001 (0.000)	−0.001 (0.000)
Words	0.001*** (0.000)	0.001*** (0.000)	0.001*** (0.000)
Intercept	0.100*** (0.010)	0.096*** (0.010)	0.096*** (0.010)
Fact-checker REs	✓	✓	✓
Year FEs	✓	✓	✓
#Statements	29,596	29,596	29,596

Table S3.10: Estimation results for mixed linear regression models predicting other-condemning emotions in the fact-checked statements. Fact-checker-specific random effects and yearly fixed effects are included. The statements with mixed veracity are excluded. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

S3.6.7 Propensity Score Matching

We conduct propensity score matching for true and false statements to check the robustness of the regression results for RQs1–3. Specifically, we match true and false statements based on the pre-defined variables, *i.e.*, *MTE*, *Words*, yearly dummies, organization-specific dummies, and author-specific dummies. With the caliper of 0.001, “common” support, and “noreplacement” settings, we discard 15,859 false statements and achieve all biases lower than 0.05 across the independent variables. The regression results based on the matched true and false samples are reported in Table S3.21. Additionally, to validate the results for RQ4, we conduct the same propensity score matching for statements mentioning political elites and statements not mentioning political elites according to the pre-defined variables: *Falsehood*, *MTE*, *Words*, yearly dummies, organization-specific dummies, and author-specific dummies. We achieve all biases lower than 0.05 by discarding 69 statements that mention political elites. We regress the matched statements again for predicting other-condemning emotions and report the results in Table S3.22. The results are robust and consistent with our findings in the main analysis.

Table S3.11: Estimation results for mixed logistic regression models predicting whether the fact-checked statements mention political elites of either political party [Column (1)], Republicans [Column (2)], or Democrats [Column (3)]. Mentions of politicians are validated according to Suppl. S3.2. Fact-checker-specific random effects and yearly fixed effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) Politician	(2) Republican	(3) Democrat
Falsehood	0.181 *** (0.052)	−0.308 *** (0.064)	0.632 *** (0.073)
MTE	−0.037 *** (0.004)	−0.033 *** (0.005)	−0.033 *** (0.006)
Falsehood × MTE	0.007 (0.004)	0.018 *** (0.005)	−0.004 (0.006)
Words	0.023 *** (0.002)	0.019 *** (0.002)	0.014 *** (0.002)
Intercept	−1.244 *** (0.134)	−2.157 *** (0.172)	−2.119 *** (0.153)
Fact-checker REs	✓	✓	✓
Year FEs	✓	✓	✓
#Statements	35,014	35,014	35,014

Table S3.12: Estimation results for mixed linear regression models predicting other-condemning emotions in the fact-checked statements. Mentions of politicians are validated according to Suppl. S3.2. Fact-checker-specific random effects and yearly fixed effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1)	(2)	(3)
Politician	0.009 *** (0.002)	0.020 *** (0.006)	0.022 * (0.010)
Republican			−0.003 (0.009)
Democrat			−0.002 (0.009)
Falsehood	0.007 ** (0.002)	0.008 ** (0.003)	0.008 ** (0.003)
Politician × Falsehood		−0.003 (0.005)	−0.004 (0.005)
MTE	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Politician × MTE		−0.001 * (0.000)	−0.001 * (0.000)
Words	0.001 *** (0.000)	0.001 *** (0.000)	0.001 *** (0.000)
Intercept	0.099 *** (0.009)	0.095 *** (0.009)	0.095 *** (0.009)
Fact-checker REs	✓	✓	✓
Year FEs	✓	✓	✓
#Statements	35,014	35,014	35,014

	(1) Politician	(2) Republican	(3) Democrat
Falsehood	0.192*** (0.054)	−0.308*** (0.065)	0.668*** (0.076)
MTE	−0.040*** (0.004)	−0.034*** (0.005)	−0.035*** (0.006)
Falsehood × MTE	0.008 (0.004)	0.018*** (0.005)	−0.003 (0.006)
Words	0.020*** (0.002)	0.019*** (0.002)	0.013*** (0.002)
Intercept	−1.278*** (0.136)	−2.111*** (0.173)	−2.158*** (0.158)
Fact-checker REs	✓	✓	✓
Year FEs	✓	✓	✓
#Statements	34,340	34,340	34,340

Table S3.13: Estimation results for mixed logistic regression models predicting whether the fact-checked statements mention political elites of either political party [Column (1)], Republicans [Column (2)], or Democrats [Column (3)]. Fact-checker-specific random effects and yearly fixed effects are included. The statements with mixed political elites are excluded. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1)	(2)	(3)
Politician	0.009*** (0.002)	0.018** (0.006)	0.018** (0.006)
Republican			−0.001 (0.004)
Falsehood	0.007** (0.002)	0.008** (0.003)	0.008** (0.003)
Politician × Falsehood		−0.002 (0.005)	−0.002 (0.005)
MTE	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Politician × MTE		−0.001* (0.000)	−0.001* (0.000)
Words	0.001*** (0.000)	0.001*** (0.000)	0.001*** (0.000)
Intercept	0.092*** (0.009)	0.089*** (0.009)	0.089*** (0.009)
Fact-checker REs	✓	✓	✓
Year FEs	✓	✓	✓
#Statements	34,340	34,340	34,340

Table S3.14: Estimation results for mixed linear regression models predicting other-condemning emotions in the fact-checked statements. Fact-checker-specific random effects and yearly fixed effects are included. The statements with mixed political elites are excluded. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table S3.15: Estimation results for mixed logistic regression models predicting whether the fact-checked statements mention political elites of either political party [Column (1)], Republicans [Column (2)], or Democrats [Column (3)]. Fact-checker-specific random effects, yearly fixed effects, and organization-specific fixed effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) Politician	(2) Republican	(3) Democrat
Falsehood	0.183*** (0.052)	−0.309*** (0.064)	0.633*** (0.073)
MTE	−0.037*** (0.004)	−0.033*** (0.005)	−0.032*** (0.006)
Falsehood × MTE	0.007 (0.004)	0.018*** (0.005)	−0.005 (0.006)
Words	0.022*** (0.002)	0.019*** (0.002)	0.015*** (0.002)
Intercept	−1.213*** (0.134)	−2.163*** (0.172)	−2.078*** (0.151)
Fact-checker REs	✓	✓	✓
Year FEs	✓	✓	✓
Org FEs	✓	✓	✓
#Statements	35,014	35,014	35,014

Table S3.16: Estimation results for mixed linear regression models predicting other-condemning emotions in the fact-checked statements. Fact-checker-specific random effects, yearly fixed effects, and organization-specific fixed effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1)	(2)	(3)
Politician	0.009*** (0.002)	0.021*** (0.006)	0.025* (0.010)
Republican			−0.005 (0.009)
Democrat			−0.003 (0.009)
Falsehood	0.007** (0.002)	0.008** (0.003)	0.008** (0.003)
Politician × Falsehood		−0.004 (0.005)	−0.004 (0.005)
MTE	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Politician × MTE		−0.001** (0.000)	−0.001** (0.000)
Words	0.001*** (0.000)	0.001*** (0.000)	0.001*** (0.000)
Intercept	0.098*** (0.009)	0.095*** (0.009)	0.095*** (0.009)
Fact-Checker REs	✓	✓	✓
Year FEs	✓	✓	✓
Org FEs	✓	✓	✓
#Statements	35,014	35,014	35,014

	(1) Politician	(2) Republican	(3) Democrat
Falsehood	0.197*** (0.053)	−0.290*** (0.065)	0.641*** (0.075)
MTE	−0.036*** (0.004)	−0.032*** (0.005)	−0.032*** (0.006)
Falsehood × MTE	0.006 (0.004)	0.017*** (0.005)	−0.005 (0.006)
Words	0.023*** (0.002)	0.019*** (0.002)	0.015*** (0.002)
Intercept	−1.746*** (0.445)	−2.549*** (0.566)	−2.559*** (0.631)
Fact-checker FEs	✓	✓	✓
Year FEs	✓	✓	✓
#Statements	34,323	33,692	33,870

Table S3.17: Estimation results for fixed logistic regression models predicting whether the fact-checked statements mention political elites of either political party [Column (1)], Republicans [Column (2)], or Democrats [Column (3)]. Fact-checker-specific fixed effects and yearly fixed effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1)	(2)	(3)
Politician	0.009*** (0.002)	0.020*** (0.006)	0.021* (0.010)
Republican			−0.001 (0.009)
Democrat			0.001 (0.009)
Falsehood	0.005* (0.002)	0.006* (0.003)	0.006* (0.003)
Politician × Falsehood		−0.003 (0.005)	−0.004 (0.005)
MTE	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Politician × MTE		−0.001** (0.000)	−0.001** (0.000)
Words	0.001*** (0.000)	0.001*** (0.000)	0.001*** (0.000)
Intercept	0.119*** (0.031)	0.115*** (0.031)	0.115*** (0.031)
Fact-checker FEs	✓	✓	✓
Year FEs	✓	✓	✓
#Statements	35,014	35,014	35,014

Table S3.18: Estimation results for fixed linear regression models predicting other-condemning emotions in the fact-checked statements. Fact-checker-specific fixed effects and yearly fixed effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table S3.19: Estimation results for mixed logistic regression models predicting whether the fact-checked statements mention political elites of either political party [Column (1)], Republicans [Column (2)], or Democrats [Column (3)]. Fact-checker-specific random effects, yearly fixed effects, and author-specific fixed effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) Politician	(2) Republican	(3) Democrat
Falsehood	0.206*** (0.053)	−0.167* (0.066)	0.526*** (0.074)
MTE	−0.037*** (0.004)	−0.031*** (0.005)	−0.036*** (0.006)
Falsehood × MTE	0.007 (0.004)	0.017*** (0.005)	−0.003 (0.006)
Words	0.022*** (0.002)	0.018*** (0.002)	0.015*** (0.002)
Intercept	−1.455*** (0.137)	−2.617*** (0.184)	−2.117*** (0.159)
Fact-Checker REs	✓	✓	✓
Year FEs	✓	✓	✓
Author FEs	✓	✓	✓
#Statements	35,014	35,014	35,014

Table S3.20: Estimation results for mixed linear regression models predicting other-condemning emotions in the fact-checked statements. Fact-checker-specific random effects, yearly fixed effects, and author-specific fixed effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1)	(2)	(3)
Politician	0.010*** (0.002)	0.022*** (0.006)	0.025* (0.010)
Republican			−0.004 (0.009)
Democrat			−0.003 (0.009)
Falsehood	0.006* (0.002)	0.006* (0.003)	0.006* (0.003)
Politician × Falsehood		−0.003 (0.005)	−0.003 (0.005)
MTE	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Politician × MTE		−0.001** (0.000)	−0.001** (0.000)
Words	0.001*** (0.000)	0.001*** (0.000)	0.001*** (0.000)
Intercept	0.109*** (0.009)	0.106*** (0.009)	0.106*** (0.009)
Fact-Checker REs	✓	✓	✓
Year FEs	✓	✓	✓
Author FEs	✓	✓	✓
#Statements	35,014	35,014	35,014

	(1) Politician	(2) Republican	(3) Democrat
Falsehood	0.317*** (0.063)	0.074 (0.077)	0.462*** (0.088)
MTE	−0.037*** (0.004)	−0.033*** (0.005)	−0.034*** (0.006)
Falsehood × MTE	0.001 (0.005)	0.001 (0.006)	0.004 (0.007)
Words	0.022*** (0.002)	0.018*** (0.003)	0.016*** (0.003)
Intercept	−1.566*** (0.160)	−2.818*** (0.208)	−1.975*** (0.184)
Fact-checker REs	✓	✓	✓
Year FEs	✓	✓	✓
Org FEs	✓	✓	✓
Author FEs	✓	✓	✓
#Statements	19,155	19,155	19,155

Table S3.21: Estimation results for mixed logistic regression models predicting whether the fact-checked statements mention political elites of either political party [Column (1)], Republicans [Column (2)], or Democrats [Column (3)]. The samples are filtered by propensity score matching. Fact-checker-specific random effects, yearly fixed effects, organization-specific fixed effects, and author-specific fixed effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1)	(2)	(3)
Politician	0.009*** (0.002)	0.021*** (0.006)	0.025* (0.010)
Republican			−0.004 (0.009)
Democrat			−0.004 (0.009)
Falsehood	0.006* (0.002)	0.006* (0.003)	0.006* (0.003)
Politician × Falsehood		−0.003 (0.005)	−0.003 (0.005)
MTE	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Politician × MTE		−0.001** (0.000)	−0.001** (0.000)
Words	0.001*** (0.000)	0.001*** (0.000)	0.001*** (0.000)
Intercept	0.110*** (0.009)	0.107*** (0.009)	0.107*** (0.009)
Fact-checker REs	✓	✓	✓
Year FEs	✓	✓	✓
Org FEs	✓	✓	✓
Author FEs	✓	✓	✓
#Statements	34,945	34,945	34,945

Table S3.22: Estimation results for mixed linear regression models predicting other-condemning emotions in the fact-checked statements. The samples are filtered by propensity score matching. Fact-checker-specific random effects, yearly fixed effects, organization-specific fixed effects, and author-specific fixed effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

S4

Supplementary Materials for Chapter 4

This chapter provides supplementary figures and tables for Chapter 4:
Figs. S4.1-S4.4 and Tables S4.1-S4.16.

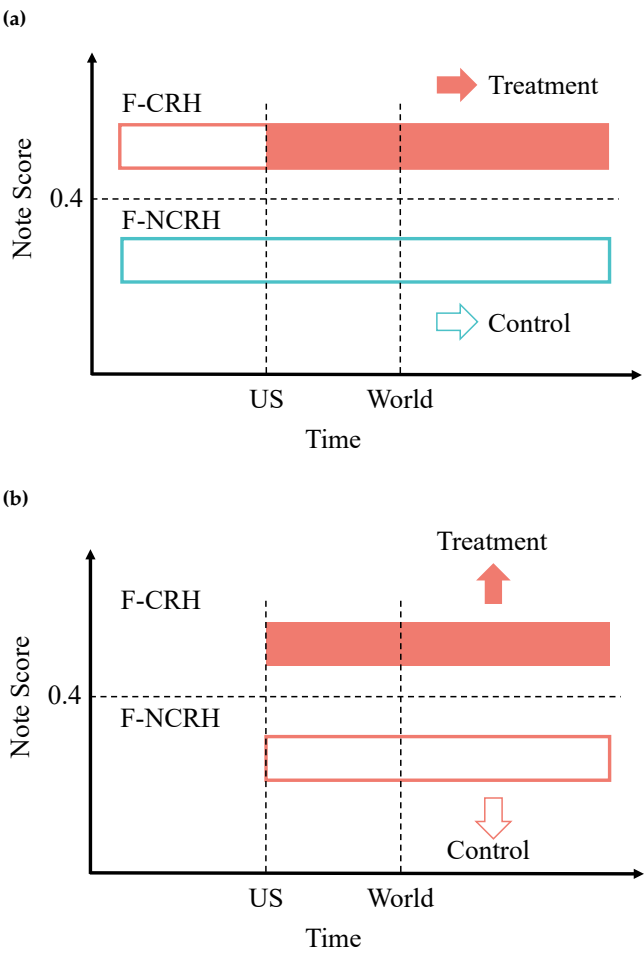


Figure S4.1: The illustrations of F-CRH (treatment) and F-NCRH (control) groups in (a) DiD and (b) RDD models.

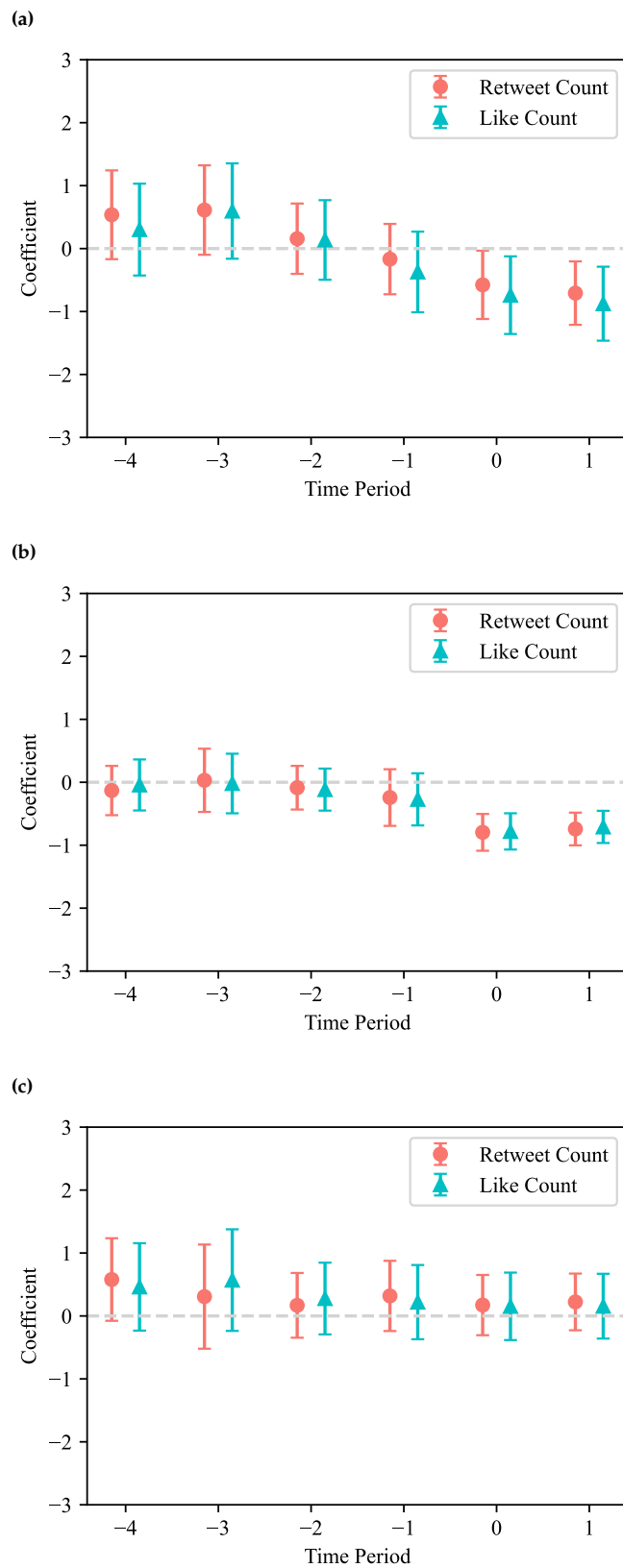


Figure S4.2: Parallel tests of retweet count and like count between the treatment and control groups. The baseline is the first three-month period spanning from July 1, 2021 to September 30, 2021. (a) Parallel test for F-CRH (treatment) and F-NCRH (control) groups. The regression results are shown in Table S4.11 and S4.12. (b) Parallel test for T-NMR (treatment) and F-NCRH (control) groups. The regression results are shown in Table S4.13 and S4.14. (c) Parallel test for F-CRH (treatment) and T-NMR (control) groups. The regression results are shown in Table S4.15 and S4.16.

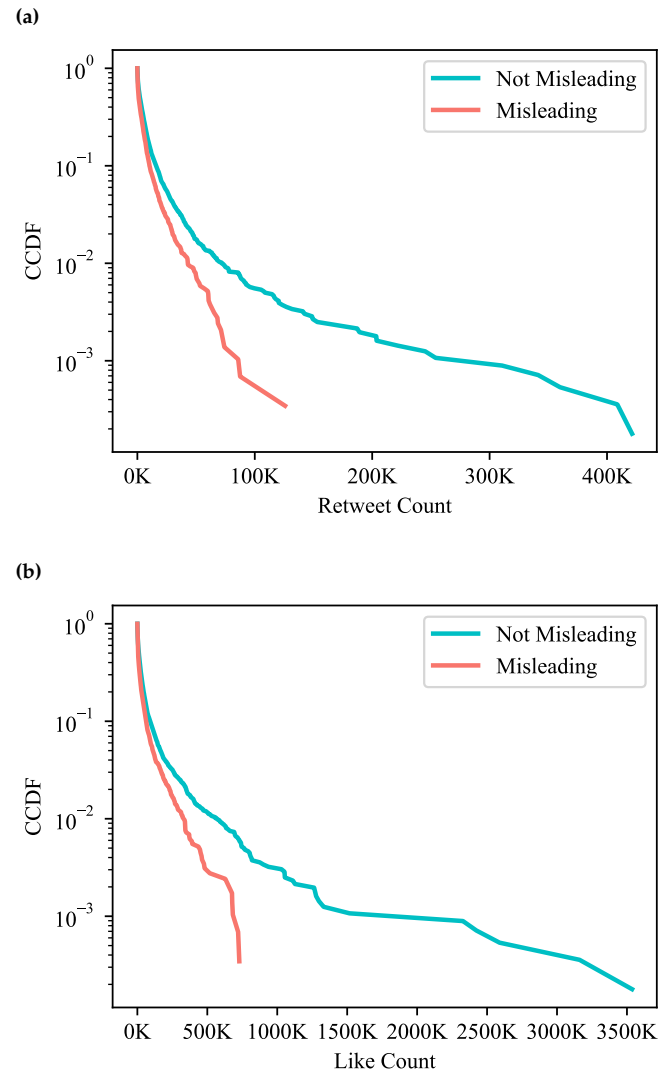


Figure S4.3: CCDFs of retweet count and like count based on the fact-checked tweets from matched F-CRH (Misleading) and T-NMR (Not Misleading) groups. (a) CCDFs of retweet count for misleading and not misleading tweets. $KS = 0.121, p < 0.001$. (b) CCDFs of like count for misleading and not misleading tweets. $KS = 0.133, p < 0.001$.

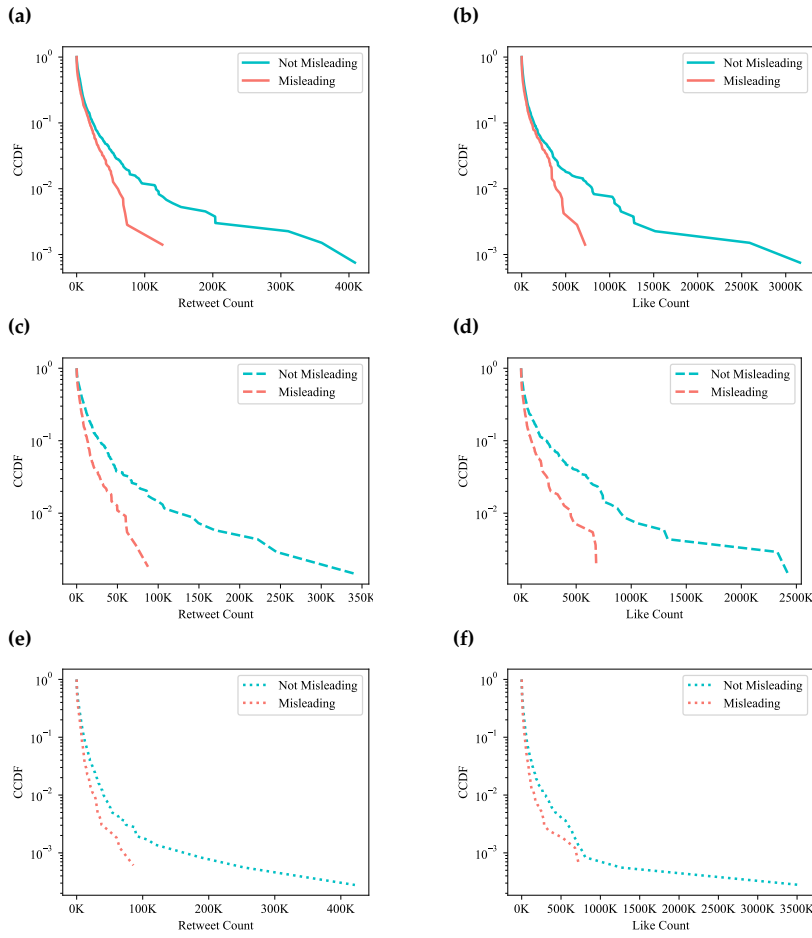


Figure S4.4: CCDFs of retweet count and like count based on the fact-checked tweets from matched F-CRH (Misleading) and T-NMR (Not Misleading) groups, during different time periods. (a) CCDFs of retweet count for misleading and not misleading tweets before 2022-10-06. $KS = 0.153, p < 0.001$. (b) CCDFs of like count for misleading and not misleading tweets before 2022-10-06. $KS = 0.168, p < 0.001$. (c) CCDFs of retweet count for misleading and not misleading tweets between 2022-10-06 and 2022-12-11 (US). $KS = 0.186, p < 0.001$. (d) CCDFs of like count for misleading and not misleading tweets between 2022-10-06 and 2022-12-11 (US). $KS = 0.205, p < 0.001$. (e) CCDFs of retweet count for misleading and not misleading tweets after 2022-12-11 (World). $KS = 0.128, p < 0.001$. (f) CCDFs of like count for misleading and not misleading tweets after 2022-12-11 (World). $KS = 0.133, p < 0.001$.

Table S4.1: Topic list with corresponding domains

Topic	Number	Domain
Entertainment	54	Actor, American Football Game, Athlete, Award Show, Baseball Game, Basketball Game, Book, Book Genre, Coach, Concert, Cricket Match, Entertainment Personality, Esports League, Esports Player, Esports Team, Exercise & Fitness, Fan Community, Fictional Character, Global TV Show, Hockey Game, Holiday, Interests and Hobbies, Interests and Hobbies Category, Interests and Hobbies Vertical, Movie, Movie Festival, Movie Genre, Multimedia Franchise, Music Album, Music Genre, Musician, NFL Football Game, Podcast, Points of Interest, Radio Station, Soccer Match, Sport, Sports Event, Sports League, Sports Personality, Sports Series, Sports Team, TV Channels, TV Episodes, TV Genres, TV Shows, Travel, Video Game, Video Game Conference, Song, Video Game Hardware, Video Game Personality, Video Game Publisher, Video Game Tournament
Finance & Business	8	Brand, Brand Vertical, Business Taxonomy, Digital Assets & Crypto, Product, Product Taxonomy, Product Version, Stocks
Politics	3	Political Body, Political Race, Politician
Science & Technology	3	Fields of study, Technology, Video Game Hardware
Emergency Events	1	Emergency Events
Ongoing News Story	1	Ongoing News Story

Table S4.2: Descriptive statistics

Variable	Count	Mean	Std	Min	25%	50%	75%	Max
RetweetCount	30267	2865.39	9402.57	0	27	346	2336.5	421330
LikeCount	30267	16470.3	66307.5	0	125	1544	10447	3.53659e+06
WordNum	30267	15.45	7.923	0	9	15	22	81
Media	30267	0.408	0.492	0	0	0	1	1
Hashtag	30267	0.098	0.298	0	0	0	0	1
Mention	30267	0.209	0.406	0	0	0	0	1
AcctAge	30267	3093.3	1790.13	0	1300.5	3577	4711	19305
Verified	30267	0.674	0.469	0	0	1	1	1
Followers	30267	3.56078e+06	1.65064e+07	0	14589	141159	746093	1.34151e+08
Followees	30267	8143.63	39808.2	0	388.5	1176	3444.5	610998
Anger	30267	0.148	0.184	0	0	0.111	0.25	1
Disgust	30267	0.078	0.142	0	0	0	0.143	1
Joy	30267	0.131	0.246	0	0	0	0.167	1
Fear	30267	0.213	0.24	0	0	0.2	0.333	1
Sadness	30267	0.129	0.174	0	0	0	0.222	1
Surprise	30267	0.108	0.196	0	0	0	0.167	1
EmergencyEvents	30267	0.001	0.024	0	0	0	0	1
Entertainment	30267	0.236	0.424	0	0	0	0	1
FinanceBusiness	30267	0.259	0.438	0	0	0	1	1
OngoingNewsStory	30267	0.143	0.35	0	0	0	0	1
Politics	30267	0.336	0.472	0	0	0	1	1
ScienceTechnology	30267	0.127	0.334	0	0	0	0	1

	(1) Retweet Count	(2) Like Count
<i>FCRH</i>	1.236*** (0.144)	1.282*** (0.152)
<i>Vis</i>	0.145 (0.080)	0.152 (0.090)
<i>FCRH</i> $\times$ <i>Vis</i>	−0.487** (0.152)	−0.455** (0.161)
<i>WordNum</i>	0.141 *** (0.028)	−0.015 (0.030)
<i>Media</i>	0.668*** (0.053)	0.512*** (0.058)
<i>Hashtag</i>	−0.742*** (0.089)	−0.986*** (0.090)
<i>Mention</i>	−1.301*** (0.066)	−1.323*** (0.069)
<i>AcctAge</i>	−0.017 (0.023)	0.009 (0.025)
<i>Verified</i>	0.764*** (0.056)	0.703*** (0.059)
<i>Followers</i>	0.180*** (0.013)	0.223*** (0.012)
<i>Followees</i>	0.029 (0.016)	0.031** (0.012)
<i>VisWorld</i>	−0.782*** (0.057)	−0.834*** (0.061)
Emotions	✓	✓
Topics	✓	✓
<i>Intercept</i>	6.847*** (0.082)	8.611 *** (0.095)
Observations	13,808	13,808

Table S4.3: DiD regression results for retweet count and like count within F-CRH (treatment) and F-NCRH (control) groups between Pilot and Roll-Out periods. Negative binomial regression models with robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

Table S4.4: DiD regression results for retweet count and like count within T-NMR (treatment) and F-NCRH (control) groups. Negative binomial regression models with robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

	Retweet Count		Like Count	
	(1) Pre-US	(2) US-World	(3) Pre-US	(4) US-World
<i>TNMR</i>	1.565*** (0.185)	1.026*** (0.096)	1.569*** (0.166)	1.086*** (0.092)
<i>Vis</i>	0.166* (0.081)	−0.765*** (0.064)	0.190* (0.086)	−0.816*** (0.070)
<i>TNMR × Vis</i>	−0.503* (0.204)	0.060 (0.112)	−0.439* (0.189)	0.078 (0.111)
<i>WordNum</i>	0.106** (0.038)	0.123*** (0.029)	−0.053 (0.041)	−0.042 (0.030)
<i>Media</i>	0.477*** (0.081)	0.639*** (0.054)	0.366*** (0.085)	0.409*** (0.058)
<i>Hashtag</i>	−0.686*** (0.149)	−0.618*** (0.095)	−0.884*** (0.142)	−0.851*** (0.091)
<i>Mention</i>	−1.249*** (0.099)	−1.183*** (0.065)	−1.218*** (0.108)	−1.161*** (0.067)
<i>AcctAge</i>	−0.026 (0.032)	−0.052* (0.024)	−0.001 (0.031)	−0.027 (0.026)
<i>Verified</i>	0.693*** (0.086)	0.858*** (0.061)	0.560*** (0.086)	0.775*** (0.064)
<i>Followers</i>	0.233*** (0.021)	0.235*** (0.016)	0.286*** (0.020)	0.291*** (0.015)
<i>Followees</i>	0.045 (0.047)	0.017 (0.016)	0.032 (0.051)	0.021 (0.012)
<i>Emotions</i>	✓	✓	✓	✓
<i>Topics</i>	✓	✓	✓	✓
<i>Intercept</i>	7.048*** (0.100)	6.892*** (0.078)	8.837*** (0.109)	8.706*** (0.086)
<i>Observations</i>	5,044	13,328	5,044	13,328

	Retweet Count		Like Count	
	(1) Pre-US	(2) US-World	(3) Pre-US	(4) US-World
<i>FCRH</i>	−0.267 (0.154)	−0.343* (0.168)	−0.289 (0.149)	−0.184 (0.173)
<i>Vis</i>	−0.190 (0.130)	−0.684*** (0.084)	−0.120 (0.128)	−0.749*** (0.079)
<i>FCRH × Vis</i>	−0.009 (0.216)	−0.141 (0.188)	0.113 (0.215)	−0.261 (0.198)
<i>WordNum</i>	−0.036 (0.051)	0.058 (0.032)	−0.216*** (0.051)	−0.132*** (0.033)
<i>Media</i>	−0.217* (0.105)	0.314*** (0.065)	−0.245* (0.109)	0.096 (0.069)
<i>Hashtag</i>	−0.055 (0.287)	−0.390** (0.132)	−0.305 (0.242)	−0.598*** (0.128)
<i>Mention</i>	−0.612*** (0.128)	−0.520*** (0.081)	−0.675*** (0.133)	−0.525*** (0.080)
<i>AcctAge</i>	−0.039 (0.044)	−0.109** (0.033)	0.011 (0.045)	−0.114*** (0.034)
<i>Verified</i>	−0.023 (0.130)	0.351*** (0.075)	−0.113 (0.125)	0.285*** (0.077)
<i>Followers</i>	0.255*** (0.032)	0.295*** (0.023)	0.279*** (0.030)	0.327*** (0.020)
<i>Followees</i>	0.008 (0.080)	−0.023 (0.012)	0.036 (0.084)	−0.009 (0.012)
<i>Emotions</i>	✓	✓	✓	✓
<i>Topics</i>	✓	✓	✓	✓
<i>Intercept</i>	9.182*** (0.178)	8.401*** (0.109)	10.952*** (0.166)	10.320*** (0.105)
<i>Observations</i>	1,293	4,823	1,716	6,463

Table S4.5: DiD regression results for retweet count and like count within F-CRH (treatment) and T-NMR (control) groups under the limitation of the top 25% delay in note display. Negative binomial regression models with robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

Table S4.6: DiD regression results for retweet count and like count between Pilot and Roll-Out periods in placebo tests. Negative binomial regression models with robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

	T-NMR – F-NCRH		F-CRH – T-NMR	
	(1) Retweet Count	(2) Like Count	(3) Retweet Count	(4) Like Count
<i>TNMR</i>	1.640*** (0.217)	1.623*** (0.188)		
<i>FCRH</i>			–0.315 (0.187)	–0.316 (0.173)
<i>Vis</i>	0.122 (0.079)	0.134 (0.091)	–0.265 (0.157)	–0.181 (0.143)
<i>TNMR × Vis</i>	–0.557* (0.223)	–0.464* (0.196)		
<i>FCRH × Vis</i>			–0.079 (0.195)	–0.037 (0.186)
<i>WordNum</i>	0.123*** (0.026)	–0.041 (0.027)	0.028 (0.028)	–0.141*** (0.032)
<i>Media</i>	0.617*** (0.050)	0.411*** (0.054)	0.239*** (0.055)	0.095 (0.061)
<i>Hashtag</i>	–0.644*** (0.088)	–0.869*** (0.085)	–0.495*** (0.112)	–0.678*** (0.116)
<i>Mention</i>	–1.194*** (0.060)	–1.189*** (0.062)	–0.594*** (0.071)	–0.642*** (0.071)
<i>AcctAge</i>	–0.049* (0.021)	–0.023 (0.023)	–0.089*** (0.026)	–0.076** (0.027)
<i>Verified</i>	0.792*** (0.055)	0.689*** (0.059)	0.207*** (0.061)	0.198** (0.064)
<i>Followers</i>	0.237*** (0.015)	0.296*** (0.014)	0.292*** (0.022)	0.328*** (0.019)
<i>Followees</i>	0.019 (0.016)	0.022 (0.012)	–0.012 (0.012)	0.005 (0.013)
<i>VisWorld</i>	–0.747*** (0.054)	–0.796*** (0.058)	–0.643*** (0.062)	–0.727*** (0.063)
Emotions	✓	✓	✓	✓
Topics	✓	✓	✓	✓
<i>Intercept</i>	6.847*** (0.081)	8.653*** (0.095)	8.837*** (0.163)	10.603*** (0.145)
Observations	15,354	15,354	6,944	6,944

	Retweet Count		Like Count	
	(1) Pre-US	(2) US-World	(3) Pre-US	(4) US-World
<i>FCRH</i>	1.172 *** (0.134)	0.860 *** (0.117)	1.224 *** (0.140)	0.947 *** (0.117)
<i>Vis</i>	0.182 (0.304)	0.019 (0.125)	0.121 (0.310)	0.198 (0.136)
<i>FCRH × Vis</i>	−0.438 ** (0.163)	−0.102 (0.132)	−0.411 * (0.168)	−0.083 (0.135)
<i>Musk</i>	0.097 (0.199)	0.165 (0.260)	0.193 (0.187)	0.213 (0.239)
<i>OldWriter</i>	0.026 (0.091)	−0.110 (0.070)	0.033 (0.093)	−0.026 (0.070)
<i>WordNum</i>	0.082 * (0.037)	0.139 *** (0.031)	−0.072 (0.041)	−0.011 (0.032)
<i>Media</i>	0.510 *** (0.076)	0.763 *** (0.059)	0.454 *** (0.083)	0.598 *** (0.059)
<i>Hashtag</i>	−0.870 *** (0.150)	−0.707 *** (0.091)	−1.059 *** (0.143)	−0.896 *** (0.094)
<i>Mention</i>	−1.184 *** (0.101)	−1.284 *** (0.071)	−1.168 *** (0.108)	−1.301 *** (0.068)
<i>AcctAge</i>	−0.017 (0.034)	−0.040 (0.024)	−0.004 (0.033)	−0.032 (0.025)
<i>Verified</i>	0.604 *** (0.082)	0.817 *** (0.061)	0.508 *** (0.083)	0.797 *** (0.060)
<i>Followers</i>	0.160 *** (0.015)	0.181 *** (0.012)	0.199 *** (0.015)	0.224 *** (0.012)
<i>Followees</i>	0.037 (0.035)	0.037 * (0.015)	0.042 (0.039)	0.044 *** (0.011)
<i>MonthYear</i>				
2022/08	0.007 (0.131)		−0.026 (0.147)	
2022/09	0.040 (0.144)		−0.077 (0.148)	
2022/10	−0.056 (0.295)		−0.181 (0.309)	
2022/11	−0.029 (0.342)	−0.038 (0.233)	−0.041 (0.350)	0.111 (0.213)
2022/12	−0.244 (0.349)	−0.296 (0.242)	−0.270 (0.360)	−0.179 (0.224)
2023/01		−0.785 ** (0.258)		−0.922 *** (0.246)
2023/02		−1.186 *** (0.263)		−1.400 *** (0.253)
2023/03		−1.229 *** (0.261)		−1.389 *** (0.249)
2023/04		−1.055 *** (0.271)		−1.355 *** (0.255)
<i>Emotions</i>	✓	✓	✓	✓
<i>Topics</i>	✓	✓	✓	✓
<i>Intercept</i>	7.024 *** (0.151)	6.941 *** (0.166)	8.816 *** (0.163)	8.490 *** (0.155)
<i>Observations</i>	5,019	11,821	5,019	11,821

Table S4.7: DiD regression results with external controls for retweet count and like count within F-CRH (treatment) and F-NCRH (control) groups. Negative binomial regression models with robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

Table S4.8: DiD regression results with external controls for retweet count plus, like count plus, and retweets including quotes within F-CRH (treatment) and F-NCRH (control) groups. Negative binomial regression models with robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

	Retweet Count Plus		Like Count Plus		Retweets + Quotes	
	(1) Pre-US	(2) US-World	(3) Pre-US	(4) US-World	(5) Pre-US	(6) US-World
<i>FCRH</i>	1.174 *** (0.133)	0.865 *** (0.114)	1.225 *** (0.140)	0.949 *** (0.117)	1.163 *** (0.128)	0.911 *** (0.120)
<i>Vis</i>	0.179 (0.300)	0.019 (0.122)	0.120 (0.310)	0.197 (0.135)	0.217 (0.288)	0.053 (0.123)
<i>FCRH × Vis</i>	−0.430 ** (0.161)	−0.053 (0.129)	−0.410 * (0.168)	−0.073 (0.135)	−0.375 * (0.160)	−0.079 (0.135)
<i>Musk</i>	0.096 (0.197)	0.155 (0.254)	0.193 (0.186)	0.211 (0.238)	0.041 (0.206)	0.104 (0.272)
<i>OldWriter</i>	0.029 (0.090)	−0.101 (0.068)	0.034 (0.093)	−0.025 (0.069)	0.019 (0.094)	−0.104 (0.068)
<i>WordNum</i>	0.081 * (0.037)	0.136 *** (0.030)	−0.072 (0.041)	−0.010 (0.032)	0.074 (0.038)	0.128 *** (0.030)
<i>Media</i>	0.505 *** (0.076)	0.740 *** (0.057)	0.454 *** (0.082)	0.595 *** (0.059)	0.583 *** (0.079)	0.790 *** (0.057)
<i>Hashtag</i>	−0.863 *** (0.148)	−0.696 *** (0.089)	−1.057 *** (0.143)	−0.890 *** (0.093)	−0.828 *** (0.155)	−0.714 *** (0.090)
<i>Mention</i>	−1.176 *** (0.100)	−1.252 *** (0.068)	−1.167 *** (0.107)	−1.294 *** (0.067)	−1.145 *** (0.104)	−1.258 *** (0.070)
<i>AcctAge</i>	−0.017 (0.034)	−0.037 (0.024)	−0.004 (0.033)	−0.031 (0.024)	−0.015 (0.035)	−0.040 (0.024)
<i>Verified</i>	0.602 *** (0.081)	0.814 *** (0.060)	0.508 *** (0.082)	0.796 *** (0.060)	0.664 *** (0.083)	0.843 *** (0.060)
<i>Followers</i>	0.160 *** (0.015)	0.180 *** (0.012)	0.199 *** (0.015)	0.224 *** (0.012)	0.161 *** (0.015)	0.188 *** (0.012)
<i>Followees</i>	0.035 (0.035)	0.036 * (0.015)	0.042 (0.039)	0.044 *** (0.011)	0.028 (0.034)	0.036 * (0.015)
<i>MonthYear</i>						
2022/08	0.007 (0.130)		−0.026 (0.147)		0.001 (0.130)	
2022/09	0.043 (0.143)		−0.076 (0.148)		0.004 (0.141)	
2022/10	−0.051 (0.292)		−0.179 (0.308)		−0.078 (0.281)	
2022/11	−0.018 (0.338)	−0.027 (0.226)	−0.039 (0.349)	0.112 (0.212)	−0.089 (0.334)	−0.073 (0.247)
2022/12	−0.229 (0.345)	−0.282 (0.235)	−0.267 (0.359)	−0.177 (0.223)	−0.334 (0.340)	−0.358 (0.254)
2023/01		−0.755 ** (0.251)		−0.916 *** (0.245)		−0.845 ** (0.270)
2023/02		−1.148 *** (0.256)		−1.392 *** (0.252)		−1.281 *** (0.273)
2023/03		−1.190 *** (0.254)		−1.381 *** (0.248)		−1.304 *** (0.272)
2023/04		−1.018 *** (0.263)		−1.347 *** (0.254)		−1.157 *** (0.281)
<i>Emotions</i>	✓	✓	✓	✓	✓	✓
<i>Topics</i>	✓	✓	✓	✓	✓	✓
<i>Intercept</i>	7.030 *** (0.150)	6.952 *** (0.165)	8.817 *** (0.163)	8.493 *** (0.154)	7.200 *** (0.148)	7.160 *** (0.162)
<i>Observations</i>	5,019	11,821	5,019	11,821	5,019	11,821

	Retweet Count		Like Count	
	(1) Pre-US	(2) US-World	(3) Pre-US	(4) US-World
<i>FCRH</i>	1.177 *** (0.136)	0.830*** (0.109)	1.220*** (0.141)	0.896*** (0.108)
<i>Vis</i>	0.076 (0.312)	0.053 (0.146)	−0.012 (0.326)	0.163 (0.165)
<i>FCRH × Vis</i>	−0.454* (0.182)	0.030 (0.148)	−0.390* (0.186)	0.120 (0.153)
<i>Musk</i>	0.236 (0.261)	0.149 (0.238)	0.257 (0.245)	0.217 (0.216)
<i>OldWriter</i>	−0.019 (0.120)	−0.115 (0.072)	0.045 (0.123)	−0.071 (0.073)
<i>WordNum</i>	0.087 (0.045)	0.120** (0.039)	−0.053 (0.048)	−0.003 (0.041)
<i>Media</i>	0.502*** (0.090)	0.699*** (0.076)	0.451*** (0.099)	0.577*** (0.074)
<i>Hashtag</i>	−0.898*** (0.155)	−0.835*** (0.138)	−1.085*** (0.152)	−1.045*** (0.131)
<i>Mention</i>	−1.198 *** (0.123)	−1.241*** (0.104)	−1.197 *** (0.132)	−1.179 *** (0.101)
<i>AcctAge</i>	0.004 (0.041)	−0.089** (0.032)	−0.001 (0.042)	−0.076* (0.032)
<i>Verified</i>	0.627*** (0.094)	0.863*** (0.081)	0.515*** (0.097)	0.796*** (0.078)
<i>Followers</i>	0.170 *** (0.031)	0.183*** (0.014)	0.232*** (0.038)	0.219*** (0.013)
<i>Followees</i>	0.043 (0.043)	0.024 (0.022)	0.052 (0.042)	0.046* (0.021)
<i>MonthYear</i>				
2022/08	−0.014 (0.130)		−0.047 (0.144)	
2022/09	0.042 (0.146)		−0.077 (0.148)	
2022/10	−0.032 (0.298)		−0.122 (0.321)	
2022/11	−0.118 (0.392)	−0.023 (0.209)	0.010 (0.410)	0.107 (0.187)
2022/12	−0.344 (0.404)	−0.291 (0.218)	−0.198 (0.425)	−0.187 (0.200)
2023/01		−0.907*** (0.259)		−0.954*** (0.253)
2023/02		−1.226*** (0.259)		−1.441 *** (0.255)
2023/03		−1.317*** (0.257)		−1.447*** (0.251)
2023/04		−1.205*** (0.295)		−1.487*** (0.277)
<i>Emotions</i>	✓	✓	✓	✓
<i>Topics</i>	✓	✓	✓	✓
<i>Intercept</i>	7.123 *** (0.177)	6.959*** (0.184)	8.875*** (0.187)	8.567*** (0.165)
<i>Observations</i>	3,814	6,021	3,814	6,021

Table S4.9: DiD regression results with external controls for retweet count and like count within F-CRH (treatment) and F-NCRH (control) matched groups during different periods. Negative binomial regression models with robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

Table S4.10: DiD regression results with external controls for retweet count and like count within F-CRH (treatment) and OCN (control) groups. Negative binomial regression models with robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

	Retweet Count		Like Count	
	(1) Pre-US	(2) US-World	(3) Pre-US	(4) US-World
<i>FCRH</i>	2.835*** (0.315)	2.464*** (0.275)	2.623*** (0.284)	2.460*** (0.254)
<i>Vis</i>	0.220 (0.552)	−0.151 (0.351)	0.144 (0.532)	0.102 (0.361)
<i>FCRH × Vis</i>	−0.669 (0.422)	−0.121 (0.305)	−0.524 (0.429)	−0.401 (0.290)
<i>Musk</i>	−0.133 (0.570)	−0.200 (0.400)	−0.230 (0.554)	−0.193 (0.416)
<i>WordNum</i>	−0.098 (0.134)	−0.091 (0.061)	−0.259 (0.136)	−0.194** (0.067)
<i>Media</i>	0.284 (0.259)	0.327** (0.115)	0.522* (0.246)	0.402*** (0.118)
<i>Hashtag</i>	−0.802* (0.403)	−1.423*** (0.169)	−0.904* (0.373)	−1.528*** (0.179)
<i>Mention</i>	−0.880* (0.352)	−0.487** (0.177)	−0.667 (0.373)	−0.571** (0.181)
<i>AcctAge</i>	0.371** (0.122)	−0.144** (0.054)	0.326** (0.122)	−0.174** (0.055)
<i>Verified</i>	0.009 (0.241)	0.572*** (0.135)	−0.075 (0.225)	0.687*** (0.137)
<i>Followers</i>	0.362*** (0.092)	0.143*** (0.022)	0.396*** (0.106)	0.156*** (0.024)
<i>Followees</i>	0.011 (0.116)	−0.044 (0.032)	0.113 (0.173)	−0.054 (0.037)
<i>MonthYear</i>				
2022/08	0.132 (0.331)		−0.184 (0.318)	
2022/09	−0.049 (0.286)		−0.032 (0.292)	
2022/10	−0.165 (0.402)		−0.376 (0.371)	
2022/11	0.481 (0.674)	0.287 (0.371)	0.574 (0.626)	0.468 (0.392)
2022/12	0.445 (0.752)	0.057 (0.401)	0.590 (0.694)	0.110 (0.421)
2023/01		−0.376 (0.471)		−0.351 (0.503)
2023/02		0.043 (0.472)		−0.001 (0.501)
2023/03		−0.127 (0.458)		−0.088 (0.493)
2023/04		−0.586 (0.498)		−0.341 (0.536)
<i>Emotions</i>	✓	✓	✓	✓
<i>Topics</i>	✓	✓	✓	✓
<i>Intercept</i>	5.952*** (0.380)	5.441*** (0.352)	7.877*** (0.351)	7.096*** (0.319)
Observations	538	1,874	538	1,874

Variable	Coef.	Robust Std. Err.	z	P> z	95%CI	
<i>FCRH</i>	1.402	0.251	5.587	0.000	0.910	1.893
<i>FCRH</i> × <i>Pre4</i>	0.537	0.360	1.493	0.136	-0.168	1.242
<i>FCRH</i> × <i>Pre3</i>	0.612	0.363	1.688	0.091	-0.099	1.322
<i>FCRH</i> × <i>Pre2</i>	0.157	0.285	0.550	0.582	-0.402	0.716
<i>FCRH</i> × <i>Pre1</i>	-0.167	0.285	-0.586	0.558	-0.726	0.392
<i>FCRH</i> × <i>VisUS</i>	-0.576	0.277	-2.083	0.037	-1.119	-0.034
<i>FCRH</i> × <i>VisWorld</i>	-0.708	0.257	-2.752	0.006	-1.212	-0.204
<i>WordNum</i>	0.119	0.022	5.365	0.000	0.076	0.163
<i>Media</i>	0.496	0.043	11.632	0.000	0.413	0.580
<i>Hashtag</i>	-0.469	0.086	-5.436	0.000	-0.638	-0.300
<i>Mention</i>	-1.271	0.058	-21.776	0.000	-1.386	-1.157
<i>AcctAge</i>	-0.066	0.020	-3.355	0.001	-0.104	-0.027
<i>Verified</i>	0.695	0.044	15.772	0.000	0.609	0.781
<i>Followers</i>	0.189	0.012	16.027	0.000	0.166	0.212
<i>Followees</i>	0.063	0.017	3.741	0.000	0.030	0.095
<i>Emotions</i>			✓			
<i>Topics</i>			✓			
<i>Periods</i>			✓			
<i>Intercept</i>	6.088	0.069	88.591	0.000	5.953	6.222
<i>Observations</i>			24,018			

Table S4.11: Regression results for retweet count parallel test between F-CRH and F-CNRH groups. Negative binomial regression model with robust standard errors.

Variable	Coef.	Robust Std. Err.	z	P> z	95%CI	
<i>FCRH</i>	1.670	0.293	5.694	0.000	1.095	2.245
<i>FCRH</i> × <i>Pre4</i>	0.302	0.372	0.811	0.417	-0.428	1.031
<i>FCRH</i> × <i>Pre3</i>	0.596	0.387	1.543	0.123	-0.161	1.354
<i>FCRH</i> × <i>Pre2</i>	0.137	0.323	0.424	0.671	-0.495	0.769
<i>FCRH</i> × <i>Pre1</i>	-0.371	0.327	-1.137	0.256	-1.012	0.269
<i>FCRH</i> × <i>VisUS</i>	-0.740	0.315	-2.352	0.019	-1.357	-0.124
<i>FCRH</i> × <i>VisWorld</i>	-0.876	0.300	-2.921	0.003	-1.463	-0.288
<i>WordNum</i>	-0.036	0.023	-1.560	0.119	-0.081	0.009
<i>Media</i>	0.334	0.046	7.208	0.000	0.243	0.425
<i>Hashtag</i>	-0.697	0.075	-9.261	0.000	-0.845	-0.550
<i>Mention</i>	-1.289	0.056	-22.974	0.000	-1.399	-1.179
<i>AcctAge</i>	-0.057	0.021	-2.711	0.007	-0.098	-0.016
<i>Verified</i>	0.611	0.047	12.964	0.000	0.518	0.703
<i>Followers</i>	0.239	0.011	21.088	0.000	0.217	0.261
<i>Followees</i>	0.054	0.013	4.135	0.000	0.028	0.080
<i>Emotions</i>			✓			
<i>Topics</i>			✓			
<i>Periods</i>			✓			
<i>Intercept</i>	7.656	0.071	108.502	0.000	7.518	7.794
<i>Observations</i>			24,018			

Table S4.12: Regression results for like count parallel test between F-CRH and F-CNRH groups. Negative binomial regression model with robust standard errors.

Variable	Coef.	Robust Std. Err.	z	P> z	95%CI	
<i>TNMR</i>	1.841	0.120	15.313	0.000	1.605	2.076
<i>TNMR</i> × <i>Pre4</i>	-0.131	0.200	-0.658	0.511	-0.523	0.260
<i>TNMR</i> × <i>Pre3</i>	0.031	0.256	0.121	0.904	-0.471	0.533
<i>TNMR</i> × <i>Pre2</i>	-0.086	0.177	-0.488	0.625	-0.433	0.260
<i>TNMR</i> × <i>Pre1</i>	-0.243	0.230	-1.058	0.290	-0.693	0.207
<i>TNMR</i> × <i>VisUS</i>	-0.796	0.149	-5.337	0.000	-1.088	-0.504
<i>TNMR</i> × <i>VisWorld</i>	-0.742	0.133	-5.601	0.000	-1.002	-0.483
<i>WordNum</i>	0.109	0.021	5.265	0.000	0.068	0.149
<i>Media</i>	0.466	0.040	11.524	0.000	0.387	0.545
<i>Hashtag</i>	-0.421	0.084	-5.029	0.000	-0.585	-0.257
<i>Mention</i>	-1.184	0.054	-22.015	0.000	-1.289	-1.078
<i>AcctAge</i>	-0.079	0.018	-4.328	0.000	-0.115	-0.043
<i>Verified</i>	0.714	0.043	16.560	0.000	0.630	0.799
<i>Followers</i>	0.249	0.014	17.797	0.000	0.221	0.276
<i>Followees</i>	0.051	0.017	2.966	0.003	0.017	0.084
<i>Emotions</i>			✓			
<i>Topics</i>			✓			
<i>Periods</i>			✓			
<i>Intercept</i>	6.083	0.067	91.447	0.000	5.953	6.214
<i>Observations</i>			26,016			

Table S4.13: Regression results for retweet count parallel test between T-NMR and F-CNRH groups. Negative binomial regression model with robust standard errors.

Table S4.14: Regression results for like count parallel test between T-NMR and F-CNRH groups. Negative binomial regression model with robust standard errors.

Variable	Coef.	Robust Std. Err.	z	P> z	95%CI	
<i>TNMR</i>	1.884	0.117	16.143	0.000	1.655	2.113
<i>TNMR</i> × <i>Pre4</i>	-0.043	0.207	-0.206	0.837	-0.448	0.363
<i>TNMR</i> × <i>Pre3</i>	-0.020	0.242	-0.082	0.935	-0.493	0.454
<i>TNMR</i> × <i>Pre2</i>	-0.117	0.170	-0.685	0.494	-0.451	0.217
<i>TNMR</i> × <i>Pre1</i>	-0.271	0.211	-1.285	0.199	-0.686	0.143
<i>TNMR</i> × <i>VisUS</i>	-0.781	0.147	-5.313	0.000	-1.069	-0.493
<i>TNMR</i> × <i>VisWorld</i>	-0.709	0.131	-5.433	0.000	-0.965	-0.453
<i>Media</i>	0.270	0.043	6.252	0.000	0.186	0.355
<i>Hashtag</i>	-0.635	0.072	-8.828	0.000	-0.775	-0.494
<i>Mention</i>	-1.195	0.052	-23.108	0.000	-1.297	-1.094
<i>AcctAge</i>	-0.069	0.020	-3.519	0.000	-0.108	-0.031
<i>Verified</i>	0.609	0.046	13.281	0.000	0.520	0.699
<i>Followers</i>	0.318	0.013	23.661	0.000	0.292	0.344
<i>Followees</i>	0.043	0.013	3.242	0.001	0.017	0.069
Emotions			✓			
Topics			✓			
Periods			✓			
<i>Intercept</i>	7.670	0.068	112.803	0.000	7.537	7.803
Observations			26,016			

Table S4.15: Regression results for retweet count parallel test between F-CRH and T-NMR groups. Negative binomial regression model with robust standard errors.

Variable	Coef.	Robust Std. Err.	z	P> z	95%CI	
<i>FCRH</i>	-0.637	0.220	-2.895	0.004	-1.068	-0.206
<i>FCRH</i> × <i>Pre4</i>	0.579	0.335	1.730	0.084	-0.077	1.235
<i>FCRH</i> × <i>Pre3</i>	0.307	0.422	0.728	0.467	-0.520	1.135
<i>FCRH</i> × <i>Pre2</i>	0.169	0.262	0.642	0.521	-0.346	0.683
<i>FCRH</i> × <i>Pre1</i>	0.318	0.285	1.119	0.263	-0.239	0.876
<i>FCRH</i> × <i>VisUS</i>	0.173	0.244	0.707	0.479	-0.306	0.651
<i>FCRH</i> × <i>VisWorld</i>	0.223	0.230	0.970	0.332	-0.227	0.673
<i>WordNum</i>	0.021	0.026	0.821	0.412	-0.029	0.071
<i>Media</i>	0.245	0.051	4.827	0.000	0.145	0.344
<i>Hashtag</i>	-0.452	0.096	-4.727	0.000	-0.639	-0.265
<i>Mention</i>	-0.587	0.065	-9.049	0.000	-0.714	-0.460
<i>AcctAge</i>	-0.074	0.023	-3.161	0.002	-0.119	-0.028
<i>Verified</i>	0.082	0.054	1.517	0.129	-0.024	0.187
<i>Followers</i>	0.291	0.020	14.574	0.000	0.252	0.330
<i>Followees</i>	-0.016	0.012	-1.334	0.182	-0.038	0.007
Emotions			✓			
Topics			✓			
Periods			✓			
<i>Intercept</i>	8.479	0.113	75.182	0.000	8.258	8.701
Observations			8,499			

Table S4.16: Regression results for like count parallel test between F-CRH and T-NMR groups. Negative binomial regression model with robust standard errors.

Variable	Coef.	Robust Std. Err.	z	P> z	95%CI	
<i>FCRH</i>	-0.527	0.251	-2.102	0.036	-1.019	-0.036
<i>FCRH</i> × <i>Pre4</i>	0.462	0.354	1.303	0.193	-0.233	1.156
<i>FCRH</i> × <i>Pre3</i>	0.570	0.412	1.383	0.167	-0.237	1.377
<i>FCRH</i> × <i>Pre2</i>	0.277	0.291	0.952	0.341	-0.293	0.847
<i>FCRH</i> × <i>Pre1</i>	0.220	0.301	0.730	0.465	-0.370	0.809
<i>FCRH</i> × <i>VisUS</i>	0.153	0.273	0.559	0.576	-0.383	0.689
<i>FCRH</i> × <i>VisWorld</i>	0.155	0.263	0.591	0.554	-0.359	0.670
<i>WordNum</i>	-0.149	0.029	-5.099	0.000	-0.206	-0.092
<i>Media</i>	0.097	0.055	1.750	0.080	-0.012	0.206
<i>Hashtag</i>	-0.586	0.101	-5.816	0.000	-0.783	-0.388
<i>Mention</i>	-0.636	0.066	-9.657	0.000	-0.765	-0.507
<i>AcctAge</i>	-0.072	0.025	-2.940	0.003	-0.121	-0.024
<i>Verified</i>	0.051	0.056	0.909	0.363	-0.059	0.161
<i>Followers</i>	0.336	0.018	19.047	0.000	0.302	0.371
<i>Followees</i>	0.004	0.013	0.313	0.754	-0.021	0.029
Emotions			✓			
Topics			✓			
Periods			✓			
<i>Intercept</i>	10.111	0.113	89.632	0.000	9.890	10.332
Observations			8,499			

S5

Supplementary Materials for Chapter 5

S5.1 Influencer Bias Chart

AllSides evaluates the media bias of top political influencers on the social media platform X (<https://www.allsides.com/media-bias/x-bias-chart>). The bias chart for political influencers is shown in Fig. S5.1.

AllSides X Influencer Bias Chart™

Ratings based on online content only. Ratings do not reflect accuracy or credibility; they reflect perspective only.



AllSides Media Bias Ratings™ are based on multi-partisan analysis. Visit [AllSides.com](https://www.allsides.com) for balanced news and over 2,400 rated sources.

Version 1.1
©AllSides 2025

Figure S5.1: Influencer bias chart.

S5.2 DiD Specification

Leads-and-lags specification. To examine temporal dynamics before and after likes were made private, we estimate a leads-and-lags Difference-in-Differences (DiD) specification:

$$\log(E(Y_{i,t})|x_{i,t}) = \beta_0 + \beta_1 D_i + \beta'_2 \text{Before}_t + \beta'_3 \text{After}_t + \beta'_4 D_i \times \text{Before}_t + \beta'_5 D_i \times \text{After}_t + \alpha_i, \quad (\text{S5.1})$$

where D_i is a dummy variable indicating whether post i belongs to the high-reputational-risk group ($= 1$) or the low-reputational-risk group ($= 0$). When comparing likes with reposts within the same posts, D_i instead indicates whether the engagement type is a like ($= 1$) or a repost ($= 0$).

The vectors Before_t and After_t denote week dummies relative to the policy implementation date. Specifically, Before_t covers weeks -8 to -2 , with week -1 serving as the reference period, while After_t spans weeks $+1$ to $+4$. Thus, β_2 and β_3 capture time fixed effects, and α_i denotes post-specific fixed effects. Finally, the interaction coefficients β_4 represent pre-treatment (lead) effects, used to assess the parallel trends assumption, whereas β_5 capture post-treatment (lag) effects, measuring the weekly impact of the private-like visibility policy on liking behavior.

Additionally, we examine the aggregated effect of the private visibility based on the following DiD specification with two periods, *i.e.*, pre- and post-policy change:

$$\log(E(Y_{i,t})|x_{i,t}) = \beta_0 + \beta_1 D_i + \beta_2 \text{After}_t + \beta_3 D_i \times \text{After}_t + \alpha_i + \gamma_t, \quad (\text{S5.2})$$

where After_t is a post-treatment indicator that equals to 1 after the implementation of private visibility and 0 before. The coefficient estimate of β_3 captures the aggregated effect of the private visibility on the liking behavior. α_i and γ_t represent post-specific and weekly-specific fixed effects, respectively.

Transformation of effect size. The coefficients estimated for the DiD terms in the two-period and multi-period regression models are the natural logarithms of the ratios for the number of likes with the treatment of private visibility compared to the number of likes that are expected to receive with public visibility during the post-treatment period. We exponentially transform the coefficient estimates of the DiD terms in the models and measure the effect size of the private visibility as:

$$\text{effect} = e^\beta - 1, \quad (\text{S5.3})$$

where β is the coefficient estimates for the DiD terms. *Effect* indicates the ratio of extra change of the number of likes to posts with the treatment of private visibility relative to the number of likes that are expected to receive with public visibility. We used this indicator to analyze the effect of private visibility on liking behavior.

S5.3 HonestDiD Estimation

Fig S5.3 presents the HonestDiD estimates for the post-change weekly periods.

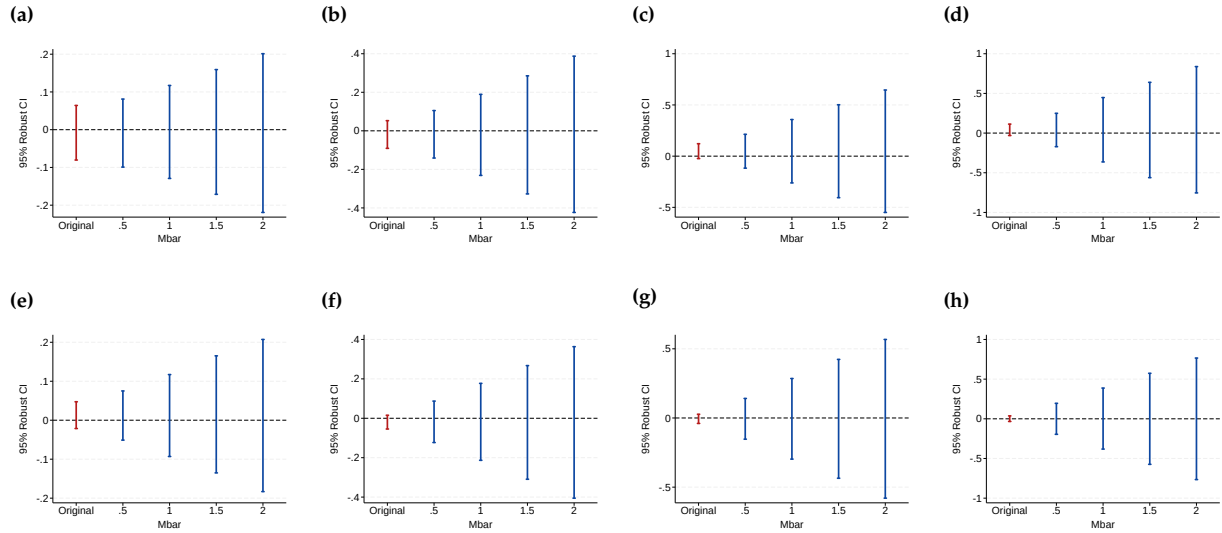


Figure S5.2: The HonestDiD estimates for the post-change weekly periods.

S5.4 Reweighting

Fig. S5.3 presents the weighted and unweighted demographics of participants. The weighted demographics align well with the demographic distribution of overall population on $\mathbb{X}$.

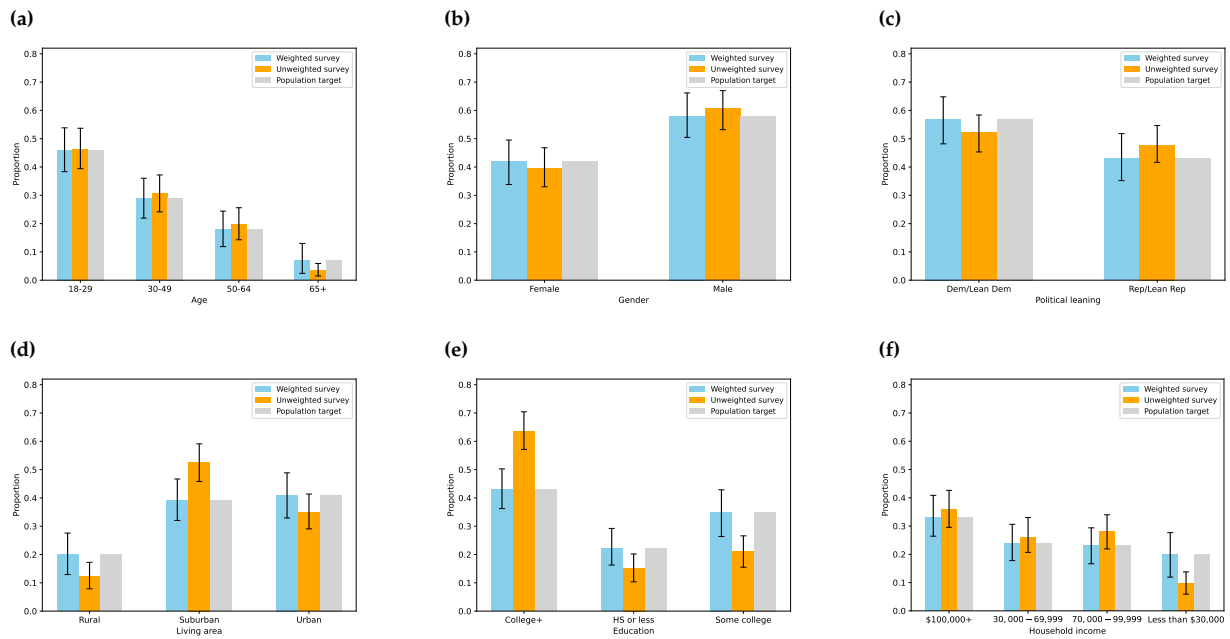


Figure S5.3: Demographics of participants. (a) The age distribution of participants. (b) The gender distribution of participants. (c) The political leaning distribution of participants. (d) The living area distribution of participant. (e) The education distribution of participants. (f) The household income distribution of participants. The error bars represent 95% CIs.

S6

Supplementary Materials for Chapter 6

S6.1 Data Overview

The “Community Notes” program on X enables users to flag posts that they believe are misleading and contribute textual notes that provide context to the source post (see example in Fig. S6.1a). After a community note is submitted, the fact-check becomes available for other enrolled contributors to rate while remaining hidden from the public users. To surface helpful community notes that appeal broadly across heterogeneous user groups, the program features a bridging-based rating system [67]. This system calculates the helpfulness score for each community note based on the ratings made by the contributors. Only notes that are rated as helpful by multiple contributors with heterogeneous rating histories are displayed to public users on X [418]. Importantly, X’s recommendation algorithm (available as open source [295]) does not impose any penalties (*e.g.*, visibility reduction) on posts flagged with community notes [296]. This implies that any observed changes in resharing behavior after a community note becomes visible must be attributed to shifts in user behavior, rather than algorithmic intervention (see Suppl. S6.11 for details).

To evaluate the efficacy of community notes in reducing the spread of misleading posts on X, we gathered time series data on repost counts and fact-checking histories for 237,180 posts that have been fact-checked on X’s “Community Notes” platform (see Section 6.4). Our dataset covers all community fact-checked posts written in English between the roll-out of “Community Notes” to the general public on October 6, 2022, and June 11, 2024, *i.e.*, for an observation period of over 20 months. For each post, we used the X API v2 to retrieve the time series of repost histories at minute level over 36 hours after its creation, during which the vast majority of all reposts occurred [10, 180]. Throughout our observation period, the posts in our dataset had been reposted more than 431 million times. Subsequently, we combined the time series data on repost counts with fact-checking histories indicating at which time point the community notes were rated helpful and displayed to the public users. The data collection process resulted in a longitudinal dataset comprising information about the repost counts at minute level and the current note status for each fact-checked post (see Suppl. S6.2 for summary statistics). This allows us to scrutinize the efficacy of community notes in reducing the spread of misleading posts on X.

Fig. S6.1b shows that the “Community Notes” feature gained significant traction after its roll-out to the public on October 6, 2022. Both the number of community notes and fact-checked source posts increased significantly. On average, community note contributors fact-checked 386 posts in total per day. Here, multiple users can write notes for the same post. Therefore, the data sometimes includes multiple fact-checks for the same post. The average number of notes per fact-checked post was

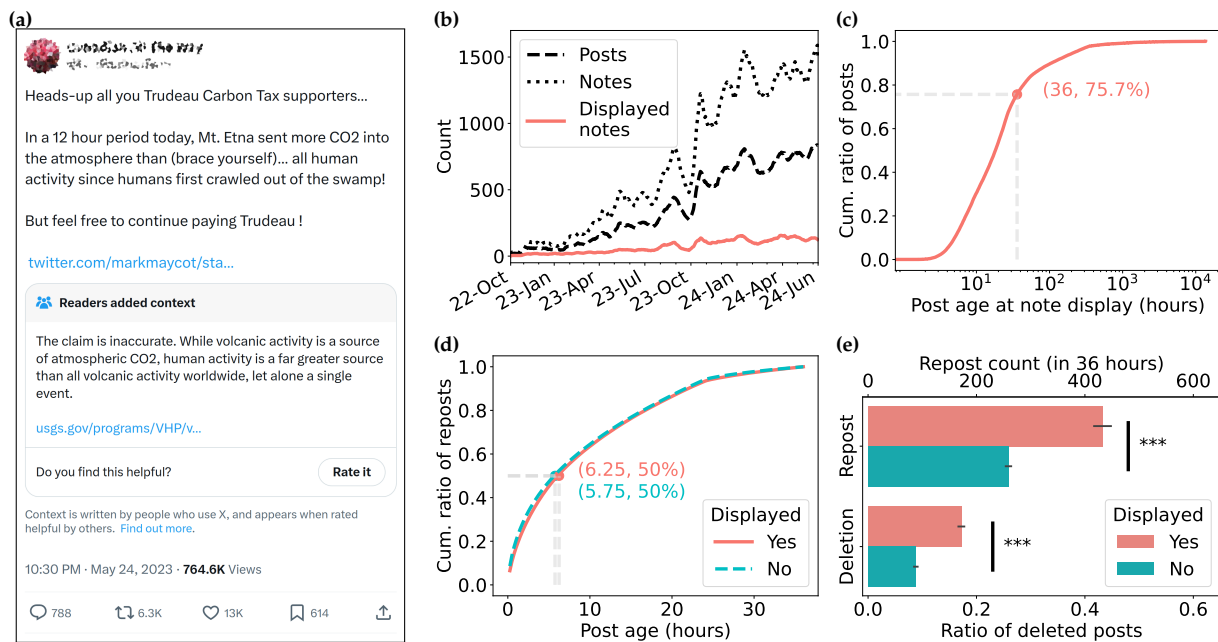


Figure S6.1: Data overview. (a) An example of a community note displayed on a misleading post on X. (b) Two-week rolling averages of the daily counts of fact-checked source posts, community notes, and displayed notes in our observation period from October 6, 2022 to June 11, 2024. (c) The cumulative distribution of the ratios of posts that received displayed notes at different post ages relative to all posts with displayed notes. (d) The cumulative distribution of the ratios of reposts at different post ages relative to all reposts within 36 hours. (e) The repost counts of retrieved posts (upper half plot, shown are median values) and the ratios of deleted posts (bottom half plot, shown are mean values) within groups of posts with displayed notes and posts without displayed notes. The error bars represent 99% Confidence Intervals (CIs).

1.867. Only a relatively small fraction of 10.1% of all notes was actually displayed to users, *i.e.*, were surfaced as helpful by the rating system of “Community Notes.” However, once community notes received the helpful status, they tended to remain displayed stably on the fact-checked posts (see Suppl. S6.2.2).

Compared to the rapid dissemination of posts on X, fact-checking via community notes was relatively slow. While 75.7% of all notes deemed helpful were displayed to general users within 36 hours after their creation, the time lag between post creation and note display (*i.e.*, the response time) averaged at 62.9 hours (median of 18.1 hours; see Fig. S6.1c). In contrast, posts on X spread considerably faster: the average half-life of posts (*i.e.*, the post age at which the cumulative ratio of reposts reaches to 50% of 36-hour reposts) amounted to merely 6.25 hours for posts with displayed notes (Fig. S6.1d). For posts without displayed notes, the average half-life of posts is 5.75 hours, which is slightly but significantly ($t = 13.135, p < 0.001$) smaller than for posts with displayed notes. Fig. S6.1e further shows that source posts with displayed notes tended to receive a higher cumulative number of 36-hour reposts (median of 432) on average, compared to those that did not have any displayed notes (median of 259, $KS = 0.117, p < 0.001$). Overall, these observations support earlier findings [74], indicating that community notes might be too slow to curb overall engagement with misleading posts on X. Notwithstanding, even if community notes might be too slow to have a meaningful effect at the cumulative level, it is still entirely possible (and plausible) that they reduced reposts for posts with displayed notes relative to similar posts without displayed notes. Quantifying this intervention effect presents the main objective of our study.

The data further suggests that posts with displayed notes were more

likely to be deleted on $\mathbb{X}$ (see Fig. S6.1e). On average, the deletion ratio was 17.3% for posts with displayed notes and 9.7% for posts without displayed notes ($t = 26.809, p < 0.001$).

S6.2 Descriptive Statistics

An overview of our dataset is reported in Table S6.1. In total, the data consists of 237,180 English source posts that were fact-checked between October 6, 2022 and June 11, 2024, *i.e.*, during an observation period of approximately 20 months. These source posts were authored by 60,815 unique users and reposted more than 431 million times at the time of data collection.

We collected several variables to control for user characteristics that could potentially affect engagement with the posts. These variables include *Verified*, which is a binary indicator showing whether the author is verified ($= 1$) or not ($= 0$). Additionally, *AccountAge* represents the age of the post account in days at the time of data collection. Furthermore, *Followers* denotes the number of followers for the post author, while *Followees* indicates the number of followees for the post author.

Given that the engagement with posts might change following the development of “Community Notes” program, we controlled for the time length from the months in which the posts were created to the roll-out of the program, *i.e.*, Months From the Roll-Out (*MFRO*). Additionally, we extracted a wide range of content characteristics that may affect engagement with the source posts. These characteristics include *Media*, a binary variable indicating whether the post contains media elements ($= 1$) or not ($= 0$). Moreover, *Words* represents the word count in the source post. To determine the number of words in the source post, we performed text preprocessing by removing user mentions from the beginning of each text and eliminating all URLs. Furthermore, we converted HTML entities to Unicode characters and applied general International Components for Unicode (ICU) transforms to further normalize the character set. We then used ICU BreakIterators to identify word boundaries for splitting the texts before counting the number of tokens containing alphanumeric characters. Additionally, we computed sentiment scores for the posts in our dataset using the Twitter-RoBERTa-base model, which has been fine-tuned for sentiment analysis on tweets and achieved a high performance compared to other models [350]. We use this pre-trained machine learning model to calculate the probabilities that a given post conveys positive emotion (*Positive*) and negative emotion (*Negative*).

Ultimately, we performed topic modeling to study heterogeneity across topics. Specifically, we employed (and validated) a supervised machine learning framework (see details in Supp. S6.3) to categorize the source tweets into predefined topics: (i) *Politics*, (ii) *Health*, (iii) *Economy*, (iv) *Science*. Posts that did not fall into one of these topic categories were categorized as *OTHER*.

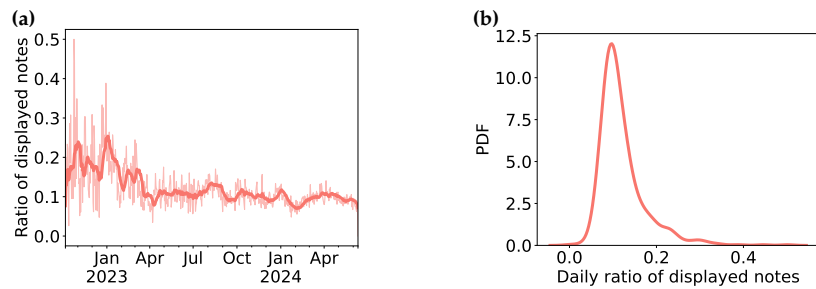
S6.2.1 Fact-Checking Activity Over Time

With the development of “Community Notes” feature on $\mathbb{X}$, the absolute number of fact-checking notes and fact-checked posts increased significantly (see Fig. S6.1b in the main paper). However, the ratio of displayed notes to all community notes over time did not increase and

Table S6.1: Dataset overview. Descriptive statistics are reported for the whole dataset (column (1)) and the subsets of posts that have received (column (2)) or have not received a displayed helpful note during our observation period (column (3)). Continuous values are reported as means with standard deviations in parentheses, unless otherwise stated.

	All	Posts with displayed notes	
		Yes	No
#Reposts	431,360,739	64,661,218	366,699,521
#Posts	237,180	36,082	201,098
#Posters	60,815	13,858	55,221
Post date	10/06/2022 – 06/11/2024	10/06/2022 – 06/10/2024	10/06/2022 – 06/11/2024
Note date	10/06/2022 – 07/13/2024	10/06/2022 – 07/11/2024	10/06/2022 – 07/13/2024
<i>Poster characteristics</i>			
Verified	75.5%	76.3%	75.3%
Account Age (days)	2957.457 (1888.851)	2730.113 (1845.222)	2998.249 (1893.691)
Followers	2,324,391 (15,269,601)	1,514,229 (10,906,328)	2,469,754 (15,922,158)
Followees	6858 (34,154)	7269 (29,461)	6784 (34,928)
<i>Post characteristics</i>			
Words	26.030 (14.888)	23.370 (14.585)	26.508 (14.891)
Media	63.6%	72.0%	62.1%
Sentiment: Positive	0.155 (0.267)	0.167 (0.275)	0.152 (0.265)
Sentiment: Negative	0.429 (0.341)	0.396 (0.338)	0.435 (0.342)
Topic: Economy	13.0%	12.3%	13.1%
Topic: Health	10.8%	9.7%	11.0%
Topic: Politics	31.9%	25.0%	33.1%
Topic: Science	11.1%	13.8%	10.7%
Topic: Other	45.7%	49.5%	45.1%
MFRO	14.236 (4.472)	13.654 (4.803)	14.340 (4.402)

Figure S6.2: The ratio of displayed notes relative to all the community notes over time. (a) Two-week rolling averages of the ratio of displayed notes over time. (b) The Probability Density Function (PDF) of daily ratio of displayed notes.



even decreased after January of 2023 (Fig. S6.2a). Notably, the ratio of displayed notes was relatively stable around 0.1 over the period after April of 2023. On average, 12.1% community notes were displayed every day, compared to the total notes that were created on the same day (Fig. S6.2b).

S6.2.2 Stability of Note Status

⌘ updates the status of notes on an hourly base according to its note ranking algorithm [418]. The rating algorithm considers all the ratings made by contributors from different perspectives to calculate the helpfulness score for each community note, determining its status as “Currently Rated Helpful,” “Needs More Ratings,” or “Currently Rated Not Helpful.” Notably, contributors can also write notes that claim the source posts are not misleading. However, such kind of notes will stay at the status of “Needs More Ratings” despite of their high helpfulness scores. Community notes that carry the status of “Currently Rated Helpful” are directly displayed on the corresponding potentially misleading posts to inform users (see an example in Fig. S6.1a). On the other hand, community notes carrying the status of “Needs More Ratings” or “Currently Rated Not

Helpful” are exclusively visible to the Community Notes contributors and, as such, are not expected to exert a substantial impact on the spread of associated source posts.

Out of the total notes, 368,910 (83.3%) remained in “Needs More Ratings (NMR)” status without any changes. 29,207 (6.6%) notes were initially rated not helpful in their first NonNMR status. Only 44,675 (10.1%) notes were initially rated helpful in their first NonNMR status with exact timestamps, and of them, 69.8% maintained helpful status in their current status and most recent status, while others changed to NMR in their current status. No note was rated helpful first and then changed into the status of not helpful in this dataset. In our main analysis, we used the first NonNMR status to indicate whether notes were displayed. Additionally, we conducted sensitivity analysis on the stability of helpful notes and found that stable helpful notes had a higher efficacy than non-stable helpful notes (see Suppl. S6.7.3).

S6.3 Identification of Topics

We used machine learning to assign topic labels to the post in our dataset. Specifically, we employed (and validated) a supervised machine learning framework to categorize the source tweets into predefined topics: (i) POLITICS, (ii) HEALTH, (iii) ECONOMY, and (iv) SCIENCE. These topics have been identified based on a manual assessment of the posts in our dataset and the selection of topics in previous works (*e.g.*, [14]). Posts that did not fall into one of these topic categories were categorized as OTHER.

Our supervised machine learning framework proceeded as follows: First, we used a pre-trained large language model as the basis for our classifier, namely, the TwHIN-BERT (large) model [419]. This model was pre-trained on a corpus of 7 billion posts from $\mathbb{X}$. Its training methodology incorporates not only text-based self-supervision (*e.g.*, masked language model) but also a social objective derived from the social engagements with posts, thereby enhancing the model’s understanding of social contexts. Second, we fine-tuned the model to our task using a manually labeled subset of posts from our dataset. Specifically, we tasked two trained research assistants to assign topic labels to a random subset of 1500 community fact-checked posts, with each assistant labeling 750 distinct posts. To ensure label accuracy, both assistants also labeled an additional 175 posts that the other had already labeled. The inter-rater reliability assessment resulted in a relatively high macro averaged Cohens’s κ coefficient of 0.711 and an overall agreement of 90.8%. Finally, we used the labeled posts to train a deep neural network classifier to predict topic labels for all posts in our dataset. All hyperparameters were tuned using 10-fold cross validation. The model was implemented in Python 3.11.3 using the Transformers Python library (version 4.30.2; [420]).

The out-of-sample prediction performance (calculated using 10-fold cross-validation) of the classifier on the manually labeled posts is reported in Table S6.2. The classifier shows a high accuracy of, on average, 0.904; and a macro-averaged F_1 score of 0.755.

Table S6.2: Out-of-sample performance in topic prediction. Performance metrics were calculated using 10-fold cross-validation.

	Accuracy	F_1
ECONOMY	0.916	0.733
HEALTH	0.944	0.845
POLITICS	0.844	0.816
SCIENCE	0.912	0.628
[Macro-]Average	0.904	0.755

S6.4 Propensity Score Matching

To reduce confounding and balance treatment and control groups, especially with respect to unmeasured confounders related to real-world events and selection bias, we performed one-to-one propensity score matching [421] to construct a control group from the source posts without displayed notes for the source posts with displayed notes. Specifically, we trained a logistic regression model based on the user-level and post-level variables (see Table S6.1) to estimate the propensity scores of all source posts. All continuous variables were z -standardized for better model fitting. According to the propensity score, we matched the closest post (caliper=0.15) in the group without displayed notes for each post in the treatment group without replacement, resulting in the creation of a matched control group for the treatment group. Additionally, we excluded matches from the source post without displayed notes in the first matching and performed the one-to-one propensity score matching again to construct another control group for the treatment group. We considered this second control group as placebo group that also shares similar user and post characteristics with the treatment group. The time when the community notes became displayed in control and placebo groups was assigned according to the corresponding posts in the treatment group. The placebo group can serve as a replacement for the treatment group in the comparison with the control group to demonstrate that there is no additional reduction in reposts without the actual treatment of note display.

As shown in Fig. S6.3, the distributions of propensity score between source posts with displayed notes and source posts without displayed notes are much more balanced after matching compared to before matching. We tested the performance of the propensity score matching using binary logistic regressions. The binary dependent variable in the logistic regressions is *Display* indicating whether the source posts have displayed community notes (= 1) or not (= 0). The explanatory variables are the user-level and post-level variables from Table S6.1. Column (1) of Table S6.3 reports the regression results for the source posts with and without displayed notes before the matching. Many of the independent variables (*e.g.*, *Verified*, *AccountAge*, *Followers*, and *Media*) are significantly different between the source posts with and without displayed notes. However, after the matching, all the independent variables among the treatment, control, and placebo groups are well-balanced and have no statistically significant differences (Columns (2) to (4) in Table S6.3).

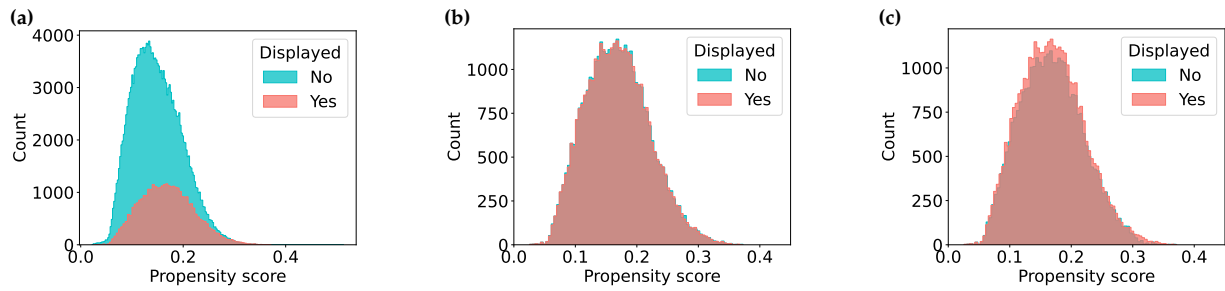


Figure S6.3: The distributions of propensity score before and after one-to-one propensity score matching. (a) The distributions of propensity score in the source posts with displayed notes and source posts without displayed notes before matching. (b) The distributions of propensity score in the source posts within the treatment group and matched source posts within the control group. (c) The distributions of propensity score in the source posts within the treatment group and matched source posts within the placebo group.

Table S6.3: Regression results for the evaluation of propensity score matching. The binary dependent variable is *Display*. Column (1) reports the results for all the source posts before matching. Column (2) reports the results for the source posts in the treatment and control groups. Column (3) reports the results for the source posts in the treatment and placebo groups. Column (4) reports the results for the source posts in the placebo and control groups. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

	(1) Before matching	(2) Treatment : Control	(3) Treatment : Placebo	(4) Placebo : Control
Verified	0.128*** (0.014)	0.010 (0.018)	-0.016 (0.018)	0.025 (0.018)
AccountAge	-0.085*** (0.006)	0.002 (0.008)	-0.003 (0.008)	0.004 (0.008)
Followers	-0.084*** (0.009)	-0.012 (0.011)	-0.017 (0.011)	0.005 (0.011)
Followees	0.026*** (0.005)	0.006 (0.008)	-0.004 (0.008)	0.008 (0.007)
Words	-0.164*** (0.006)	0.005 (0.008)	-0.006 (0.008)	0.012 (0.008)
Media	0.358*** (0.013)	0.007 (0.017)	0.020 (0.018)	-0.013 (0.018)
Economy	0.016 (0.018)	-0.016 (0.023)	-0.005 (0.024)	-0.010 (0.024)
Health	-0.096*** (0.020)	-0.010 (0.026)	0.018 (0.026)	-0.028 (0.026)
Politics	-0.261*** (0.014)	-0.008 (0.019)	-0.005 (0.019)	-0.004 (0.019)
Science	0.188*** (0.018)	-0.035 (0.023)	-0.004 (0.023)	-0.032 (0.023)
Positive	-0.023*** (0.007)	0.006 (0.009)	-0.001 (0.009)	0.007 (0.009)
Negative	-0.041*** (0.007)	0.000 (0.010)	0.015 (0.010)	-0.015 (0.010)
MFRO	-0.186*** (0.006)	0.000 (0.007)	-0.001 (0.007)	0.001 (0.008)
Intercept	-2.025*** (0.016)	-0.002 (0.021)	0.095*** (0.021)	-0.096*** (0.021)
#Posts	237,175	72,154	68,825	68,825

S6.5 Estimation Results

S6.5.1 Two-Period ATTs

The regression results for the two-period ATT estimation are reported in Table S6.4. The before-display period is between -4 and -1 hours from the note display. The period between 1 and 12 hours from the note display was considered as the after-display period.

The results for the main regression are reported in Column (1) of Table S6.4. Furthermore, Column (2) reports the results for an extended regression model that incorporates user and post characteristics as additional explanatory variables. The coefficient estimate for the difference-in-difference term ($Display \times After$) is qualitatively identical across both models, which further validates the effectiveness of the propensity score matching.

S6.5.2 Parallel Test and Multi-Period ATTs

In this section, we present the estimation results of our multi-period DiD analysis that allows us to examine the treatment effect over time. We begin by assessing parallel trends to validate the applicability of the DiD framework. Subsequently, we perform a comprehensive series of robustness checks and sensitivity analyses to further confirm the reliability of our estimates.

Estimation results and parallel trends

The regression results for the parallel test and multi-period ATTs are reported in Table S6.5. The first hour preceding the display of community notes is the baseline period in the multi-period model. The coefficient estimates of the difference-in-differences terms for the three hourly before-display periods are used to assess parallel trends relative to the baseline period. The coefficient estimates of the difference-in-differences terms in the hourly after-display periods are used to estimate hourly multi-period ATTs relative to the baseline period.

The parallel trends assumption in DiD models requires that, in the absence of treatment, the untreated potential outcomes of the treatment and control groups would follow the same trajectory. A common approach to testing this assumption is to set a null hypothesis that no pre-treatment differences exist and interpret failure to reject the null hypothesis as evidence that parallel trends hold. However, prior research suggests this approach can be misleading, especially in large datasets where statistical power is high [291–293, 422]. In such cases, even small pre-treatment differences can appear statistically significant, leading to the (incorrect) conclusion that the DiD framework is not applicable. To address this, the DiD literature proposes “non-inferiority” approaches, which reverse the null hypothesis: instead of testing whether pre-treatment differences are exactly zero, it tests whether these differences exceed a meaningful threshold [291–293]. If the data provide strong evidence that pre-trends are small, the null hypothesis of a “large pre-treatment difference” is rejected, supporting the validity of the DiD design. In the following, we (i) first assess parallel trends using conventional null hypothesis significance testing. Subsequently, we (ii) then apply equivalence testing to determine whether any detected pre-trends are practically meaningful. In addition,

we (iii) use alternative baseline periods to evaluate the sensitivity of parallel trends.

Null hypothesis testing: The coefficient estimates for $Display \times Before : 4$ ($coef. = 0.040, p < 0.001$), $Display \times Before : 3$ ($coef. = 0.043, p < 0.001$) are statistically significant at the 1% statistical significance level, whereas the coefficient estimate for $Display \times Before : 2$ ($coef. = 0.026, p = 0.011$) is statistically significant at the 5% statistical significance level. This indicates the presence of pre-trend differences between treatment and control groups. However, compared to the post-treatment effects, the effect sizes are small, making it unclear whether these differences are meaningful or merely a consequence of high statistical power. To determine whether these pre-trends are practically negligible, we conduct equivalence testing in the following.

Equivalence testing: We follow the DiD literature [291, 292] by considering an alternative null hypothesis of “non-negligible trend differences” and conducting a test for statistical equivalence by specifying a lower and upper bound. If the differences fall within the equivalence bounds, they are considered equivalent to the absence of pre-trends that are worthwhile to examine. Accordingly, we performed an equivalence test to assess whether the estimated trend differences between treatment and control groups during before-display period were large enough to be considered meaningful. As suggested in [423], the differences of 0.3 scale point on a 5-point scale can be considered as a reasonable equivalence bound. Given this, we define the bound for our equivalence test as:

$$\Delta = 0.3 * \delta_{max} / 5, \quad (S6.1)$$

where δ_{max} is the max absolute value of the estimated coefficients of difference-in-differences terms for hourly periods before and after the display of community notes. In our data, the value of δ_{max} is 1.148, and the equivalence bound is 0.069.

As shown in Fig. S6.4, all the DiD coefficient estimates during before-display period fall within the range of equivalence bounds. This means that, while statistically significant, the small trend differences during before-display period between treatment and control groups are negligible and equivalent to the absence of pre-trends. In sum, this implies that the DiD framework is applicable in our study.

Parallel test with alternative baseline periods: Our main analysis uses the hour immediately preceding note display as the baseline. To assess the sensitivity of our findings to this choice, we re-estimated the DiD model using alternative baselines, shifting the reference period to the second, third, and fourth before-display hours. Table S6.6 reports DiD coefficient estimates based on baseline periods from the second hour to the fourth hour preceding the display of community notes. In Column (1) of Table S6.6, the baseline period is the second hour before note display, and all the DiD coefficient estimates during before-display period are not statistically significant. Additionally, in Columns (2) and (3) of Table S6.6, only the coefficient estimates of $Display \times Before : 1$ are statistically significant but fall within the range of equivalence bounds. This suggests that the trend differences between treatment and control groups are too small to be considered meaningful regardless of baseline periods.

Table S6.4: Regression results for two-period ATT estimation. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

	(1) Main	(2) Extended
Display	1.200*** (0.023)	1.161*** (0.022)
After	0.106*** (0.004)	0.105*** (0.004)
Display×After	−0.948*** (0.004)	−0.948*** (0.004)
PostAge	−0.797*** (0.002)	−0.796*** (0.002)
Verified		1.144*** (0.027)
AccountAge		−0.110*** (0.012)
Followers		0.326*** (0.017)
Followees		0.116*** (0.011)
Words		0.075*** (0.012)
Media		0.774*** (0.026)
Economy		−0.368*** (0.034)
Health		0.297*** (0.038)
Politics		0.198*** (0.027)
Science		−0.581*** (0.033)
Positive		−0.094*** (0.014)
Negative		0.188*** (0.014)
MFRO		−0.180*** (0.011)
Intercept	1.293*** (0.017)	−0.071 (0.033)
Post-level RE	✓	✓
#Observations	654,400	654,400
#Posts	40,900	40,900

Figure S6.4: Equivalence test for parallel trend. Shown are hourly DiD coefficient estimates with 99% CIs. The two grey horizontal dash lines indicates the lower and upper bounds for the equivalence test.

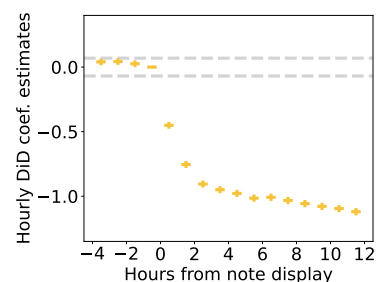


Table S6.5: Regression results for parallel test and multi-period ATTs estimation. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

Multi-period		Continued	
Display	1.175*** (0.024)	Display×Before:4	0.040*** (0.010)
Before:4	0.161*** (0.010)	Display×Before:3	0.043*** (0.010)
Before:3	0.055*** (0.009)	Display×Before:2	0.026 (0.010)
Before:2	0.014 (0.008)	Display×After:1	−0.451*** (0.010)
After:1	−0.010 (0.008)	Display×After:2	−0.754*** (0.010)
After:2	−0.031*** (0.009)	Display×After:3	−0.906*** (0.011)
After:3	−0.037*** (0.010)	Display×After:4	−0.949*** (0.011)
After:4	−0.046*** (0.011)	Display×After:5	−0.978*** (0.011)
After:5	−0.060*** (0.013)	Display×After:6	−1.015*** (0.011)
After:6	−0.050*** (0.015)	Display×After:7	−1.008*** (0.011)
After:7	−0.069*** (0.016)	Display×After:8	−1.033*** (0.011)
After:8	−0.055** (0.018)	Display×After:9	−1.057*** (0.011)
After:9	−0.046 (0.020)	Display×After:10	−1.079*** (0.011)
After:10	−0.038 (0.022)	Display×After:11	−1.096*** (0.011)
After:11	−0.026 (0.024)	Display×After:12	−1.120*** (0.011)
After:12	−0.006 (0.026)	PostAge	−0.603*** (0.015)

Continued	
Intercept	1.397*** (0.019)
Post-level RE	✓
#Observations	654,400
#Posts	40,900

Sensitivity analyses for DiD estimates

In addition to demonstrating that the pre-trends in our dataset are negligible, we followed best practices [292] and conducted a wide range of sensitivity analyses to further assess the robustness of our DiD estimates. These include (i) using bounds on relative magnitudes, (ii) incorporation of linear trend, and (iii) rolling regressions. In all of these checks, the findings consistently confirm the reliability of our estimates.

Bounds on relative magnitudes: To further assess the robustness of our estimates against potential violations of parallel trends, we employed HonestDiD using the Conditional Least Favorable Hybrid (C-LF) method (see [292, 294] for details). HonestDiD is a robust inference approach designed to quantify uncertainty in DiD estimates when strict parallel trends between treatment and control groups may not hold. Instead of assuming that any detected pre-trend invalidates the analysis, HonestDiD allows for small trend differences and adjusts the estimates accordingly. Specifically, the method assumes that any deviations from parallel trends in the after-display period cannot exceed $\bar{M}$ times the largest deviation observed in the before-display period. As shown in Fig. S6.5, our DiD

Table S6.6: Regression results parallel trends test with alternative baseline periods. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

	Base period: -2	Base period: -3	Base period: -4
Display×Before:4	0.015 (0.010)	-0.002 (0.010)	
Display×Before:3	0.017 (0.010)		0.002 (0.010)
Display×Before:2		-0.017 (0.010)	-0.015 (0.010)
Display×Before:1	-0.026 (0.010)	-0.043*** (0.010)	-0.040*** (0.010)
Display×After:1	-0.477*** (0.010)	-0.494*** (0.010)	-0.491*** (0.010)
Display×After:2	-0.780*** (0.010)	-0.797*** (0.010)	-0.795*** (0.010)
Display×After:3	-0.932*** (0.011)	-0.949*** (0.011)	-0.946*** (0.011)
Display×After:4	-0.975*** (0.011)	-0.992*** (0.011)	-0.990*** (0.011)
Display×After:5	-1.004*** (0.011)	-1.021*** (0.011)	-1.019*** (0.011)
Display×After:6	-1.041*** (0.011)	-1.058*** (0.011)	-1.056*** (0.011)
Display×After:7	-1.034*** (0.011)	-1.051*** (0.011)	-1.049*** (0.011)
Display×After:8	-1.059*** (0.011)	-1.076*** (0.011)	-1.073*** (0.011)
Display×After:9	-1.083*** (0.011)	-1.100*** (0.011)	-1.098*** (0.011)
Display×After:10	-1.105*** (0.011)	-1.122*** (0.011)	-1.119*** (0.011)
Display×After:11	-1.122*** (0.011)	-1.139*** (0.011)	-1.136*** (0.011)
Display×After:12	-1.145*** (0.011)	-1.162*** (0.011)	-1.160*** (0.011)

estimates remain robust throughout the 12 hours following the display of community notes, even when allowing after-display deviations to be twice as large as the maximum before-display deviation (*i.e.*, $\overline{M} = 2$). These results confirm that our findings hold even under relaxed parallel trends assumptions, reinforcing the credibility of our estimates.

Incorporation of linear trend: To further account for potential pre-trends, we incorporated a linear trend adjustment into the DiD model, assuming that existing trends in the before-display period continue into the after-display period [291, 422]. Based on the estimation results with varying baseline periods (see Table S6.6), we found that the pre-trend in the hour immediately before note display slightly but significantly differed from that in the three preceding hours, while the pre-trends in those three preceding hours were parallel with each other. Given this, we used the slope between the second and first before-display hours as a linear pre-trend in an augmented DiD model. The full estimation results are reported in Table S6.7. The coefficient estimate of the slope is not statistically significant (*coef.* = -0.026, $p = 0.011$). Additionally, Fig. S6.6a shows that the hourly DiD coefficient estimates between the

augmented DiD model with linear trend and the standard DiD model are not statistically significantly different from each other. Overall, this suggests that our results were robust to potential linear trends extending into the after-display period.

Rolling regressions: We conducted a series of rolling DiD regressions using a two-hour sliding window, covering the period from four hours before to twelve hours after note display. In each rolling DiD regression, we examined the difference between the treatment and control groups in the subsequent hour, relative to the previous hour. The results indicate a sharp drop in the rolling DiD coefficient immediately after note display, followed by a gradual increase that stabilizes by the fourth hour after note display (see Fig. S6.6b). This suggests that the display of community notes significantly reduced the spread of misleading posts, with its effect stabilizing within four hours.

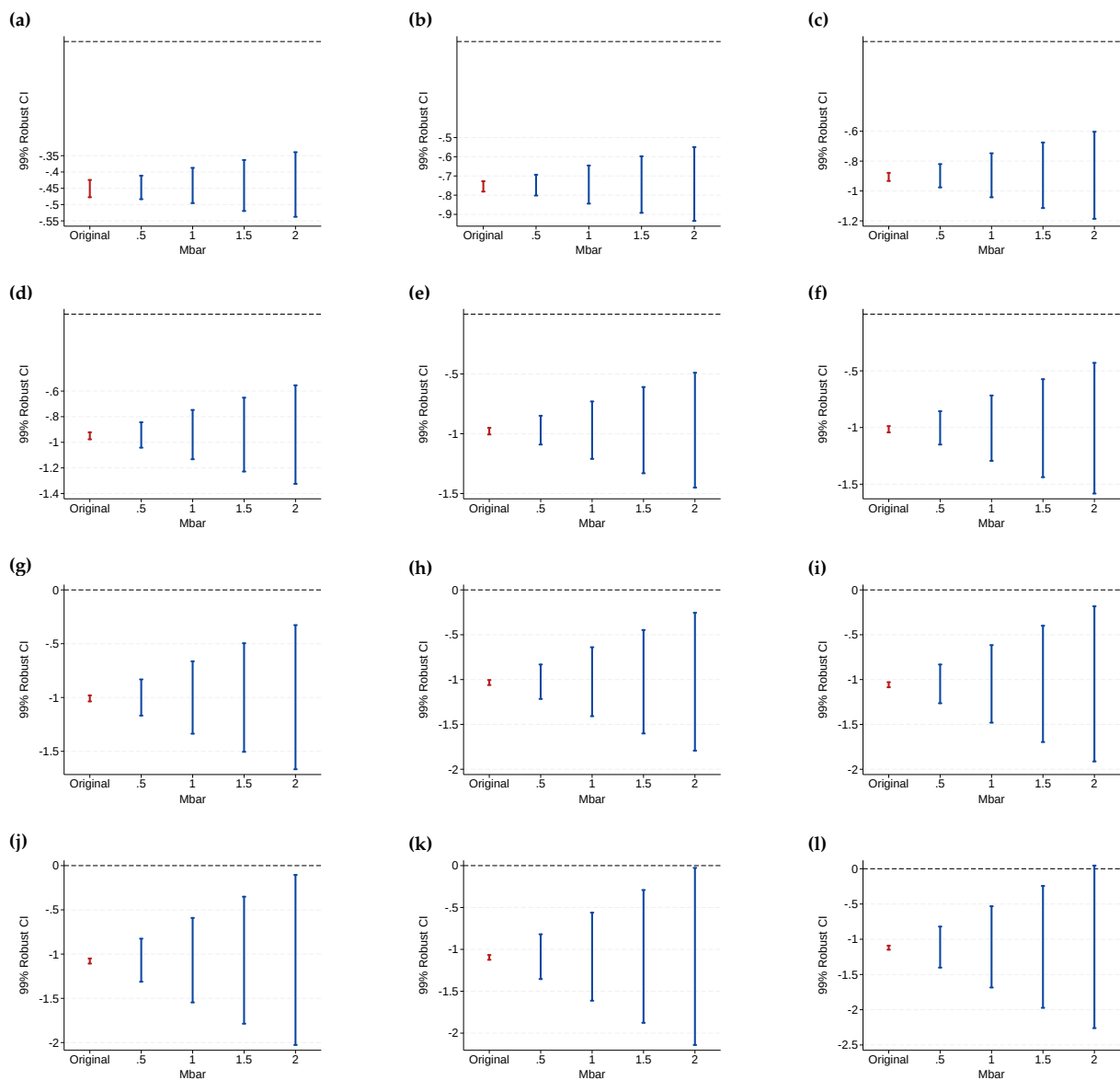


Figure S6.5: The original DiD coefficient estimates and robust confidence intervals under different $\overline{Mbar}$ from 0.5 to 2 with an increment of 0.5. (a)–(l) The first hour to the twelfth hour after the display of community notes. The error bars represent 99% CIs. The robust confidence intervals are interpreted based on HonestDiD with the C-LF method and relative magnitudes restrictions [294].

Figure S6.6: The hourly DiD coefficient estimates. (a) The hourly DiD coefficient estimates in an augmented DiD model with a linear trend. (b) The hourly DiD coefficient estimates using a rolling two-hour sliding window. The horizontal grey lines represent the upper and lower equivalence bounds. The error bars represent 99% CIs.

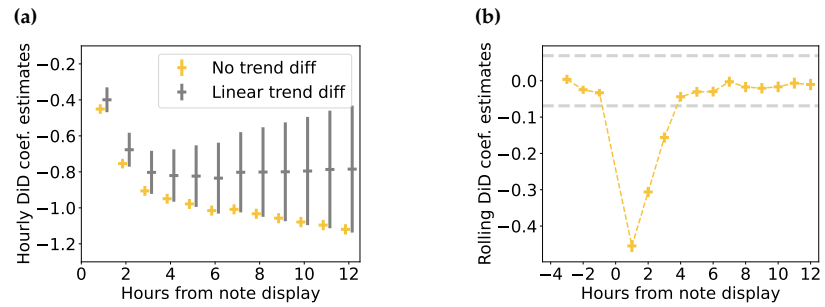


Table S6.7: Regression results for parallel test and multi-period ATTs estimation. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

Multi-period		Continued	
Display	1.149*** (0.028)	Display×Before:4	−0.037 (0.027)
Before:4	0.161*** (0.010)	Display×Before:3	−0.009 (0.017)
Before:3	0.055*** (0.009)	Display×After:1	−0.399*** (0.027)
Before:2	0.014 (0.008)	Display×After:2	−0.677*** (0.037)
After:1	−0.010 (0.008)	Display×After:3	−0.803*** (0.046)
After:2	−0.031*** (0.009)	Display×After:4	−0.821*** (0.056)
After:3	−0.037*** (0.010)	Display×After:5	−0.824*** (0.066)
After:4	−0.046*** (0.011)	Display×After:6	−0.835*** (0.076)
After:5	−0.060*** (0.013)	Display×After:7	−0.802*** (0.086)
After:6	−0.050*** (0.015)	Display×After:8	−0.801*** (0.096)
After:7	−0.069*** (0.016)	Display×After:9	−0.800*** (0.106)
After:8	−0.055** (0.018)	Display×After:10	−0.795*** (0.117)
After:9	−0.046 (0.020)	Display×After:11	−0.787*** (0.127)
After:10	−0.038 (0.022)	Display×After:12	−0.785*** (0.137)
After:11	−0.026 (0.024)	DiD Slope	−0.026 (0.010)
After:12	−0.006 (0.026)	PostAge	−0.603*** (0.015)

Continued	
Intercept	1.397*** (0.019)
Post-level RE	✓
#Observations	654,400
#Posts	40,900

S6.6 Placebo Analyses

We performed one-to-one matching again for the source posts in the treatment group and constructed a placebo group from source posts without displayed notes based on their propensity scores (see Suppl. S6.4). The placebo group demonstrates no statistically significant difference with the treatment group across the variables from user profiles and post features (Table S6.3). We considered the placebo group as a virtual treatment group that is similar to the treatment group but did not

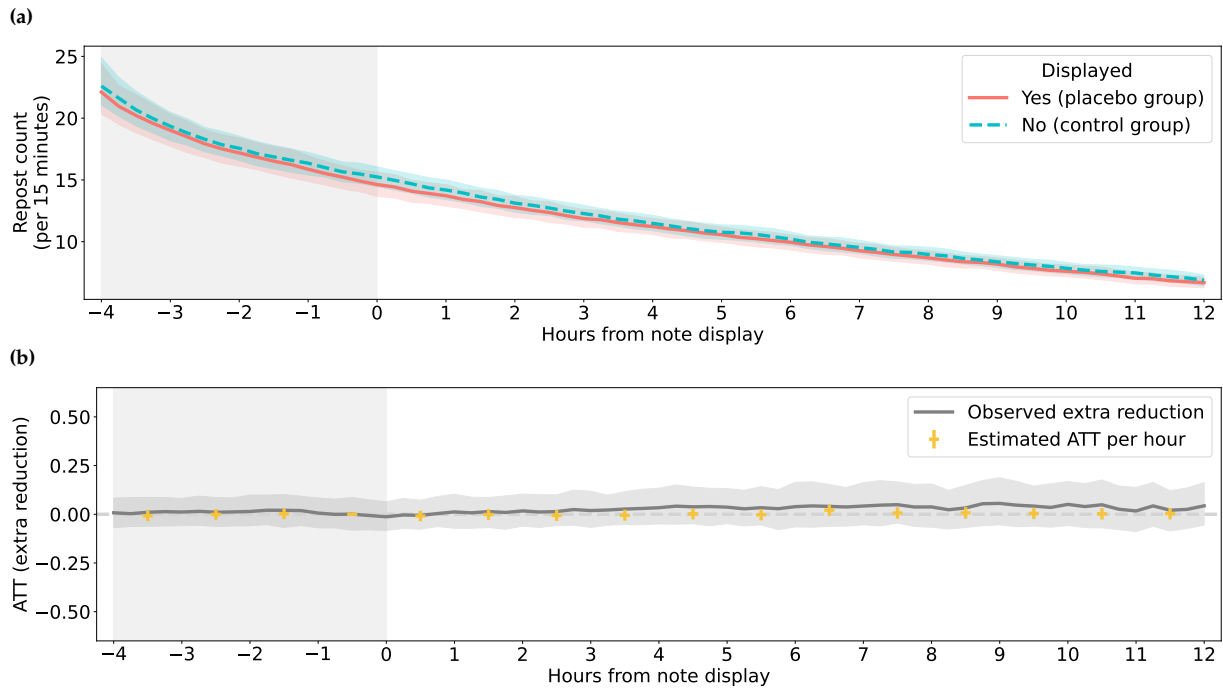


Figure S6.7: Placebo analyses for the effectiveness of community notes. (a) The changes in repost count within the source posts in the placebo and control groups. The error bands represent 99% CIs. (b) The estimated hourly extra reductions (ATTs) of reposts in placebo group after the display of community notes, compared to control group and relative to before-display period (yellow). The error bars represent 99% CIs. The grey belt (with 99% CIs) visualizes the observed extra reduction of the ratio of reposts in the treatment group relative to reposts in the control group and compared to the ratio of reposts before the display of community notes.

receive actual treatment (*i.e.*, note display). Subsequently, we investigated whether there was an additional reduction in reposts for placebo group during the after-display period compared to control group and relative to the before-display period. As shown in Fig. S6.7a, the changes in reposts within the placebo group overlapped with those within the control group. Based on the results from the regression that estimates multi-period ATTs using placebo and control groups (see Table S6.8), we found that all the hourly ATT estimations, whether before or after the note display, were not statistically significant. As visualized in Fig. S6.7b, there is no significant extra reductions after the display of community notes. Additionally, the two-period ATT estimation with the replacement of the actual treatment group was not statistically significant (ATT: 0.003; 99% CI: $[-0.009, 0.015]$; $p = 0.506$; see Table S6.9). These results suggest that the spread of the source posts in placebo and control groups followed a parallel pattern regardless of the note display.

Table S6.8: Regression results for the multi-period ATT estimation with the replacement of treatment group. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

Multi-period		Continued	
Display	−0.056 (0.027)	Display×Before:4	−0.008 (0.011)
Before:4	0.131 *** (0.010)	Display×Before:3	−0.001 (0.011)
Before:3	0.035 *** (0.009)	Display×Before:2	0.002 (0.011)
Before:2	0.004 (0.008)	Display×After:1	−0.009 (0.011)
After:1	0.002 (0.008)	Display×After:2	−0.001 (0.011)
After:2	−0.009 (0.009)	Display×After:3	−0.006 (0.011)
After:3	−0.004 (0.010)	Display×After:4	−0.005 (0.011)
After:4	−0.003 (0.012)	Display×After:5	0.002 (0.011)
After:5	−0.007 (0.014)	Display×After:6	−0.002 (0.011)
After:6	0.013 (0.016)	Display×After:7	0.021 (0.011)
After:7	0.004 (0.018)	Display×After:8	0.007 (0.011)
After:8	0.028 (0.020)	Display×After:9	0.007 (0.011)
After:9	0.047 (0.022)	Display×After:10	0.005 (0.011)
After:10	0.065* (0.024)	Display×After:11	0.002 (0.011)
After:11	0.087 *** (0.026)	Display×After:12	0.003 (0.011)
After:12	0.117 *** (0.029)	PostAge	−0.677 *** (0.017)

Continued	
Intercept	1.313 *** (0.022)
Post-level RE	✓
#Observations	623,920
#Posts	38,995

	(1) Main	(2) Extended
Display	−0.058 (0.027)	−0.057 (0.025)
After	−0.085*** (0.004)	−0.085*** (0.004)
Display×After	0.003 (0.005)	0.003 (0.005)
PostAge	−0.615*** (0.003)	−0.615*** (0.002)
Verified		1.420*** (0.030)
AccountAge		−0.065*** (0.013)
Followers		0.377*** (0.018)
Followees		0.075*** (0.012)
Words		0.109*** (0.014)
Media		0.872*** (0.028)
Economy		−0.448*** (0.038)
Health		0.321*** (0.043)
Politics		0.252*** (0.031)
Science		−0.747*** (0.038)
Positive		−0.128*** (0.015)
Negative		0.264*** (0.016)
MFRO		−0.076*** (0.012)
Intercept	1.410*** (0.018)	−0.185*** (0.036)
Post-level RE	✓	✓
#Observations	623,920	623,920
#Posts	38,995	38,995

Table S6.9: Regression results for two-period ATT estimation with the replacement of treatment group. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

S6.7 Sensitivity Analyses

We conducted additional sensitivity analyses on how the efficacy of community notes varied (i) depending on the post age at note display, (ii) over time since the roll-out of the “Community Notes” program in October 2022, (iii) depending on the the number of ratings, the stability and helpfulness scores of community notes, and (iv) across user and post characteristics.

S6.7.1 Sensitivity Across Response Time

We analyzed how the efficacy of community notes varies depending on the post age at note display, *i.e.*, the response time. We conducted the regressions across the source posts grouped by the response time with a window of 4 hours. For instance, the source posts that received displayed notes at 4 and 8 hours from post creation are grouped together for the subset analysis. The regression results are reported in Table S6.10. We found that the treatment effects were significantly more pronounced for early vs. late community notes. For instance, for community notes displayed within 4 to 8 hours since the creation of the misleading post, the ATT was -0.665 (99% CI: $[-0.671, -0.658]$; $p < 0.001$). In contrast, the ATT decreased to -0.457 (99% CI: $[-0.473, -0.439]$; $p < 0.001$) if the community notes was displayed within 20 to 24 hours after post creation. Overall, the efficacy of community notes in reducing the spread of misleading posts was higher if they were displayed earlier.

Table S6.10: Regression results for two-period ATT estimations over response time, *i.e.*, post age at note display from 4 to 24 hours. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

Response time	4 – 8	8 – 12	12 – 16	16 – 20	20 – 24
Display	1.182*** (0.040)	1.144*** (0.049)	1.134*** (0.058)	1.292*** (0.061)	1.300*** (0.065)
After	−0.142*** (0.007)	0.052*** (0.008)	0.327*** (0.010)	0.681*** (0.010)	0.114*** (0.011)
Display × After	−1.092*** (0.008)	−0.984*** (0.009)	−0.931*** (0.010)	−0.815*** (0.011)	−0.610*** (0.012)
PostAge	−0.656*** (0.004)	−0.485*** (0.005)	−0.749*** (0.006)	−1.407*** (0.007)	−1.385*** (0.007)
Intercept	1.487*** (0.029)	1.396*** (0.035)	1.279*** (0.041)	1.369*** (0.043)	1.936*** (0.045)
Post-level RE	✓	✓	✓	✓	✓
#Observations	210,112	145,408	108,832	98,464	91,584

S6.7.2 Sensitivity Across Months From Roll-Out

Next, we examined the changes in the efficacy of community notes since the roll-out of the feature in October 2022. The regression results for the two-period ATT estimation model across months from the roll-out (MFRO) of “Community Notes” program are reported in Table S6.11. We found that the ATTs in the first month and second month of the community notes program were -0.403 (99% CI: $[-0.477, -0.318]$; $p < 0.001$) and -0.322 (99% CI: $[-0.392, -0.245]$; $p < 0.001$), respectively. However, after 21 months, the ATT increased to -0.697 (99% CI: $[-0.716,$

−0.676]; $p < 0.001$). Overall, this suggests an trend of increasing efficacy of community notes in the months following its launch. This may be partly attributed to improvements within the Community Notes program, such as faster note generation (see Suppl. S6.11).

S6.7.3 Sensitivity Across Rating Thresholds and Note Helpfulness Scores

We analyzed whether the efficacy of community notes varies depending on the number of ratings, *i.e.*, the popularity of notes as reflected by other fact-checking contributors. A higher number of ratings points towards notes that attract a lot of attention from other fact-checking contributors (*e.g.*, because the fact-checked post is particularly viral). Additionally, the Community Notes system sends “Need Your Help” alerts to contributors, requesting ratings for notes with “Needs More Ratings” status. This mechanism promotes engagement with selected notes. Notably, while community notes must receive a minimum number of ratings (*e.g.*, 5) to be assigned a status, highly rated notes are not necessarily displayed, as the algorithm does not rely on majority votes to prevent manipulation. During estimation, we only considered source posts with community notes that indicate potentially misleading information and had never been rated as not helpful, as those notes were more likely to be promoted by “Needs Your Help” alerts.

Fig. S6.8 shows the ATT estimates across the rating thresholds from 10 to 80 (mean of ratings per note) with a step size of 10. We found that the ATT estimates across the rating thresholds had no significant differences with each other. This suggests that once the number of helpfulness ratings exceeded the display threshold, the efficacy of community notes in reducing the spread of misleading posts was relatively stable.

Additionally, we examined whether the stability of the helpful status affect the efficacy of community notes. Based on the original status history dataset, we categorize helpful notes into stable helpful notes (notes that retained a helpful status since their first NonNMR status) and non-stable helpful note (notes whose helpful status later disappeared). The full estimation results are reported in Table S6.13. We found that the stable helpful notes ATT: −0.654; 99% CI: [−0.658, −0.650]; $p < 0.001$) had a higher efficacy than non-stable helpful notes (ATT: −0.436; 99% CI: [−0.445, −0.426]; $p < 0.001$).

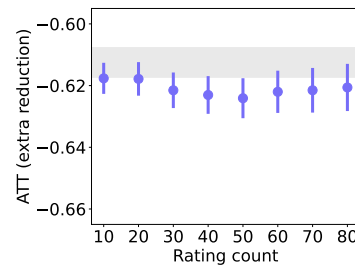
Furthermore, we analyzed how the efficacy of community notes varied across their helpfulness scores. To this end, we ran the note ranking algorithm using datasets downloaded from the Community Notes website and calculated note helpfulness scores. The full estimation results are reported in Table S6.14. We found that the efficacy of community notes increased with the increase of helpfulness scores.

Taken together, community notes reduced the spread of misleading posts across subsets separated by the number of ratings, note stability, and note helpfulness scores. Moreover, their efficacy was strengthened when having a stable helpful status and high helpfulness scores.

S6.7.4 Sensitivity Across Poster and Post Characteristics

We studied whether the treatment effect of community notes was moderated by author characteristics (*e.g.*, the number of followers, verified

Figure S6.8: The estimated ATTs of community notes depending on the number of ratings. The error bars represent 99% CIs, and the grey bands visualize the ATT (with 99% CI) estimated via the two-period DiD model from our main analysis. The full estimation results are reported in Table S6.12.



status) and post characteristics (*e.g.*, sentiment, topics). Table S6.15 reports the regression results for the sensitivity analysis on how the efficacy of community notes varied depending on poster characteristics (*Verified*, *AccountAge*, *Followers*, and *Followees*). Table S6.16 and Table S6.17 report the regression results for the sensitivity analysis on how the efficacy of community notes varied depending on post characteristics (*Words*, *Media*, *Positive*, *Negative*, *Economy*, *Health*, *Politics*, and *Science*).

We found that the ATT estimates were significantly negative across all author characteristics. However, the efficacy of community notes was higher if the author of the fact-checked post was not verified (ATT: -0.732 ; 99% CI: $[-0.741, -0.723]$; $p < 0.001$), for younger accounts (ATT: -0.651 ; 99% CI: $[-0.657, -0.645]$; $p < 0.001$), and for accounts with a lower number of followers (ATT: -0.687 ; 99% CI: $[-0.693, -0.681]$; $p < 0.001$), as compared to fact-checked post from verified accounts (ATT: -0.580 ; 99% CI: $[-0.586, -0.575]$; $p < 0.001$), older accounts (ATT: -0.570 ; 99% CI: $[-0.577, -0.563]$; $p < 0.001$), and accounts with higher number of followers (ATT: -0.549 ; 99% CI: $[-0.555, -0.543]$; $p < 0.001$). This suggests that the effectiveness of community notes might be slightly discounted for source posts from accounts with high social influence.

In terms of post characteristics, the ATT estimates were **larger** for posts that were shorter (ATT: -0.644 ; 99% CI: $[-0.650, -0.638]$; $p < 0.001$) and posts that had media elements (ATT: -0.634 ; 99% CI: $[-0.639, -0.629]$; $p < 0.001$), as compared to posts that were longer (ATT: -0.579 ; 99% CI: $[-0.585, -0.572]$; $p < 0.001$) and posts without media elements (ATT: -0.534 ; 99% CI: $[-0.545, -0.523]$; $p < 0.001$). Moreover, we used a state-of-the-art machine learning model to calculate positive and negative sentiment scores for the fact-checked posts (see Suppl. S6.2). Here, we found that the efficacy of community notes was robust and had no significant difference between posts with low negative sentiment (ATT: -0.615 ; 99% CI: $[-0.622, -0.609]$; $p < 0.001$) and those with high negative sentiment (ATT: -0.610 ; 99% CI: $[-0.616, -0.604]$; $p < 0.001$), and between posts with low positive sentiment (ATT: -0.606 ; 99% CI: $[-0.612, -0.600]$; $p < 0.001$) and those with high positive sentiment (ATT: -0.620 ; 99% CI: $[-0.626, -0.613]$; $p < 0.001$). Ultimately, we implemented (and validated) a topic modeling approach (see Suppl. S6.3) to study how the efficacy of community notes varies across different topics, namely, *Economy*, *Health*, *Politics*, and *Science*. Posts that did not fall into one of these topic categories were categorized as *OTHER*. Here, we found that the ATT was lower for posts related to economy (ATT: -0.575 ; 99% CI: $[-0.590, -0.560]$; $p < 0.001$), health (ATT: -0.549 ; 99% CI: $[-0.564, -0.535]$; $p < 0.001$), and politics (ATT: -0.567 ; 99% CI: $[-0.577, -0.558]$; $p < 0.001$), as compared to posts not related to economy (ATT: -0.617 ; 99% CI: $[-0.621, -0.612]$; $p < 0.001$), health (ATT: -0.618 ; 99% CI: $[-0.623, -0.614]$; $p < 0.001$), and politics (ATT: -0.629 ; 99% CI: $[-0.634, -0.624]$; $p < 0.001$). This suggests that users have a higher resistance towards community notes if they are attached to posts covering

economy, health, and political topics. We found no such statistically significant difference for science-related posts (ATT: -0.614 ; 99% CI: $[-0.626, -0.601]$; $p < 0.001$), as compared to posts not covering science-related topics (ATT: -0.629 ; 99% CI: $[-0.634, -0.624]$; $p < 0.001$).

Table S6.11: Regression results for two-period ATT estimations across months from the roll-out of “Community Notes” program. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

MFRO	1	2	3	4	5	6	7
Display	1.014** (0.358)	1.812*** (0.273)	0.856** (0.273)	1.663*** (0.266)	1.379*** (0.216)	0.937*** (0.169)	1.434*** (0.146)
After	0.098 (0.046)	−0.009 (0.038)	0.106* (0.041)	0.205*** (0.038)	0.192*** (0.033)	0.149*** (0.026)	0.058 (0.024)
Display × After	−0.516*** (0.051)	−0.389*** (0.042)	−0.645*** (0.042)	−0.748*** (0.042)	−0.798*** (0.036)	−0.737*** (0.029)	−0.627*** (0.026)
PostAge	−0.726*** (0.028)	−0.745*** (0.023)	−0.779*** (0.023)	−0.687*** (0.022)	−0.798*** (0.019)	−0.824*** (0.016)	−0.835*** (0.014)
Intercept	2.114*** (0.230)	1.775*** (0.181)	2.022*** (0.204)	1.585*** (0.171)	1.285*** (0.141)	1.542*** (0.109)	1.247*** (0.096)
Post-level RE	✓	✓	✓	✓	✓	✓	✓
#Observations	3568	5616	5584	6080	9136	12,992	16,304

MFRO	8	9	10	11	12	13	14
Display	1.296*** (0.124)	1.198*** (0.123)	1.667*** (0.122)	1.471*** (0.098)	1.407*** (0.108)	1.233*** (0.084)	1.017*** (0.089)
After	0.145*** (0.021)	0.156*** (0.020)	0.244*** (0.021)	0.371*** (0.017)	0.249*** (0.019)	0.084*** (0.014)	0.100*** (0.015)
Display × After	−0.663*** (0.023)	−0.640*** (0.022)	−0.631*** (0.023)	−0.800*** (0.018)	−0.714*** (0.020)	−0.967*** (0.015)	−0.957*** (0.016)
PostAge	−0.893*** (0.012)	−0.886*** (0.012)	−0.918*** (0.012)	−0.927*** (0.010)	−0.905*** (0.011)	−0.853*** (0.008)	−0.837*** (0.009)
Intercept	1.111*** (0.086)	1.349*** (0.086)	1.236*** (0.084)	1.128*** (0.070)	1.194*** (0.074)	1.411*** (0.060)	1.427*** (0.062)
Post-level RE	✓	✓	✓	✓	✓	✓	✓
#Observations	21,936	22,448	23,856	37,472	27,552	51,616	47,312

MFRO	15	16	17	18	19	20	21
Display	0.990*** (0.080)	1.222*** (0.081)	1.147*** (0.091)	1.240*** (0.074)	1.203*** (0.072)	1.197*** (0.075)	0.974*** (0.122)
After	0.072*** (0.014)	0.007 (0.014)	0.066*** (0.016)	0.069*** (0.015)	0.056*** (0.015)	0.047** (0.016)	0.021 (0.023)
Display × After	−1.014*** (0.015)	−0.989*** (0.016)	−1.089*** (0.017)	−1.070*** (0.015)	−1.066*** (0.016)	−1.188*** (0.016)	−1.192*** (0.025)
PostAge	−0.707*** (0.008)	−0.732*** (0.009)	−0.742*** (0.009)	−0.749*** (0.009)	−0.777*** (0.009)	−0.723*** (0.009)	−0.701*** (0.014)
Intercept	1.325*** (0.058)	1.316*** (0.056)	1.440*** (0.064)	1.083*** (0.056)	1.144*** (0.054)	1.158*** (0.055)	1.507*** (0.086)
Post-level RE	✓	✓	✓	✓	✓	✓	✓
#Observations	56,992	58,624	46,544	60,016	61,136	57,056	22,560

Table S6.12: Regression results for two-period ATT estimations across rating thresholds. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

Ratings	10	20	30	40	50	60	70	80
Display	0.151 *** (0.025)	-0.233 *** (0.027)	-0.414 *** (0.029)	-0.503 *** (0.031)	-0.545 *** (0.033)	-0.558 *** (0.034)	-0.564 *** (0.036)	-0.566 *** (0.038)
After	0.151 *** (0.005)	0.167 *** (0.005)	0.183 *** (0.006)	0.191 *** (0.006)	0.199 *** (0.007)	0.199 *** (0.007)	0.205 *** (0.007)	0.207 *** (0.008)
Display $\times$ After	-0.961 *** (0.005)	-0.962 *** (0.006)	-0.972 *** (0.006)	-0.976 *** (0.006)	-0.978 *** (0.007)	-0.973 *** (0.007)	-0.972 *** (0.007)	-0.969 *** (0.008)
PostAge	-0.826 *** (0.003)	-0.842 *** (0.003)	-0.849 *** (0.003)	-0.851 *** (0.003)	-0.851 *** (0.003)	-0.851 *** (0.003)	-0.848 *** (0.003)	-0.841 *** (0.003)
Intercept	2.328 *** (0.021)	2.731 *** (0.023)	2.970 *** (0.025)	3.138 *** (0.027)	3.266 *** (0.029)	3.357 *** (0.031)	3.448 *** (0.032)	3.524 *** (0.034)
Post-level RE	✓	✓	✓	✓	✓	✓	✓	✓
#Observations	495,664	448,816	413,888	382,736	351,584	324,768	298,624	272,864

	Non-stable helpful	Stable helpful
Display	1.317 *** (0.041)	1.161 *** (0.026)
After	-0.070 *** (0.005)	0.132 *** (0.004)
Display $\times$ After	-0.572 *** (0.007)	-1.061 *** (0.005)
PostAge	-0.630 *** (0.003)	-0.822 *** (0.003)
Intercept	1.412 *** (0.018)	1.267 *** (0.017)
Post-level RE	✓	✓
#Observations	403,168	578,432

Table S6.13: Regression results for two-period ATT estimations depending on the stability of note helpful status. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.**Table S6.14:** Regression results for two-period ATT estimations across note helpfulness scores. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

Note scores	0.4 – 0.45	0.45 – 0.50	0.50 – 0.55	0.55 – 0.60
Display	0.228 *** (0.044)	0.380 *** (0.040)	0.674 *** (0.043)	1.184 *** (0.070)
After	-0.010 (0.005)	0.016 ** (0.005)	0.023 *** (0.005)	-0.035 *** (0.005)
Display $\times$ After	-0.992 *** (0.008)	-1.110 *** (0.008)	-1.217 *** (0.008)	-1.109 *** (0.012)
PostAge	-0.678 *** (0.003)	-0.703 *** (0.003)	-0.709 *** (0.003)	-0.654 *** (0.004)
Intercept	2.134 *** (0.019)	2.114 *** (0.019)	2.110 *** (0.019)	2.152 *** (0.019)
Post-level RE	✓	✓	✓	✓
#Observations	287,488	298,384	288,880	252,144

Table S6.15: Regression results for two-period ATT estimations across user characteristics, *i.e.*, *Verified*, *AccountAge*, *Followers* and *Followees*. The estimation models incorporate interaction terms and subsets separately. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

	Interaction	Verified	Not verified		Interaction	High account age	Low account age
Display	1.849*** (0.049)	0.974*** (0.025)	1.877*** (0.054)	Display	1.175*** (0.024)	1.026*** (0.032)	1.377*** (0.034)
After	0.237*** (0.008)	0.085*** (0.004)	0.205*** (0.012)	After	0.102*** (0.004)	0.059*** (0.006)	0.155*** (0.006)
Display×After	−1.328*** (0.011)	−0.869*** (0.005)	−1.317*** (0.013)	Display×After	−0.933*** (0.004)	−0.843*** (0.006)	−1.053*** (0.006)
PostAge	−0.797*** (0.002)	−0.799*** (0.003)	−0.785*** (0.007)	PostAge	−0.797*** (0.002)	−0.798*** (0.003)	−0.797*** (0.004)
Verified	1.650*** (0.038)			AccountAge	0.088*** (0.017)		
Display×Verified	−0.871*** (0.055)			Display×AccountAge	−0.225*** (0.024)		
After×Verified	−0.158*** (0.009)			After×AccountAge	−0.055*** (0.003)		
ATT×Verified	0.460*** (0.012)			ATT×AccountAge	0.134*** (0.005)		
Intercept	0.028 (0.033)	1.679*** (0.018)	0.006 (0.037)	Intercept	1.301*** (0.017)	1.356*** (0.023)	1.228*** (0.024)
Post-level RE	✓	✓	✓	Post-level RE	✓	✓	✓
#Observations	654,400	508,640	145,760	#Observations	654,400	327,184	327,216

	Interaction	High followers	Low followers		Interaction	High followees	Low followees
Display	1.196*** (0.023)	0.559*** (0.029)	1.664*** (0.033)	Display	1.194*** (0.023)	1.161*** (0.032)	1.228*** (0.034)
After	0.106*** (0.004)	0.072*** (0.005)	0.148*** (0.007)	After	0.106*** (0.004)	0.116*** (0.006)	0.095*** (0.006)
Display×After	−0.943*** (0.004)	−0.796*** (0.005)	−1.162*** (0.008)	Display×After	−0.946*** (0.004)	−0.993*** (0.006)	−0.897*** (0.006)
PostAge	−0.797*** (0.002)	−0.805*** (0.003)	−0.781*** (0.004)	PostAge	−0.797*** (0.002)	−0.814*** (0.003)	−0.779*** (0.003)
Followers	0.365*** (0.024)			Followees	0.130*** (0.016)		
Display×Followers	−0.155*** (0.036)			Display×Followees	0.076** (0.024)		
After×Followers	−0.030*** (0.004)			After×Followees	0.000 (0.003)		
ATT×Followers	0.116*** (0.006)			ATT×Followees	−0.034*** (0.004)		
Intercept	1.310*** (0.017)	2.418*** (0.022)	0.292*** (0.023)	Intercept	1.293*** (0.017)	1.444*** (0.023)	1.146*** (0.024)
Post-level RE	✓	✓	✓	Post-level RE	✓	✓	✓
#Observations	654,400	327,120	327,280	#Observations	654,400	327,152	327,248

Table S6.16: Regression results for two-period ATT estimations across post characteristics including *Words*, *Media*, *Positive* and *Negative*. The estimation models incorporate interaction terms and subsets separately. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

	Interaction	High words	Low words		Interaction	Media	No media
Display	1.168*** (0.024)	1.018*** (0.032)	1.389*** (0.034)	Display	1.092*** (0.044)	1.213*** (0.027)	1.107*** (0.047)
After	0.103*** (0.004)	0.058*** (0.006)	0.159*** (0.006)	After	0.054*** (0.007)	0.131*** (0.005)	0.035*** (0.009)
Display×After	−0.933*** (0.005)	−0.864*** (0.006)	−1.034*** (0.007)	Display×After	−0.766*** (0.009)	−1.006*** (0.005)	−0.764*** (0.009)
PostAge	−0.797*** (0.002)	−0.779*** (0.003)	−0.816*** (0.004)	PostAge	−0.797*** (0.002)	−0.802*** (0.003)	−0.782*** (0.005)
Words	0.205*** (0.017)			Media	0.612*** (0.036)		
Display×Words	−0.206*** (0.024)			Display×Media	0.126 (0.052)		
After×Words	−0.032*** (0.003)			After×Media	0.070*** (0.008)		
ATT×Words	0.098*** (0.004)			ATT×Media	−0.240*** (0.010)		
Intercept	1.325*** (0.017)	1.466*** (0.023)	1.112*** (0.024)	Intercept	0.855*** (0.031)	1.469*** (0.019)	0.846*** (0.033)
Post-level RE	✓	✓	✓	Post-level RE	✓	✓	✓
#Observations	654,400	325,856	328,544	#Observations	654,400	475,008	179,392

	Interaction	High positive	Low positive		Interaction	High negative	Low negative
Display	1.189*** (0.023)	1.422*** (0.033)	0.975*** (0.033)	Display	1.184*** (0.023)	0.975*** (0.032)	1.425*** (0.033)
After	0.107*** (0.004)	0.121*** (0.006)	0.093*** (0.006)	After	0.106*** (0.004)	0.095*** (0.006)	0.118*** (0.006)
Display×After	−0.948*** (0.004)	−0.967*** (0.007)	−0.931*** (0.006)	Display×After	−0.948*** (0.004)	−0.941*** (0.006)	−0.955*** (0.007)
PostAge	−0.797*** (0.002)	−0.789*** (0.004)	−0.804*** (0.003)	PostAge	−0.797*** (0.002)	−0.794*** (0.003)	−0.800*** (0.004)
Positive	−0.372*** (0.016)			Negative	0.381*** (0.016)		
Display×Positive	0.203*** (0.023)			Display×Negative	−0.239*** (0.023)		
After×Positive	0.012*** (0.003)			After×Negative	−0.008 (0.003)		
ATT×Positive	−0.001 (0.005)			ATT×Negative	0.006 (0.005)		
Intercept	1.304*** (0.017)	0.960*** (0.024)	1.630*** (0.023)	Intercept	1.321*** (0.017)	1.649*** (0.023)	0.937*** (0.024)
Post-level RE	✓	✓	✓	Post-level RE	✓	✓	✓
#Observations	654,400	327,200	327,200	#Observations	654,400	327,200	327,200

Table S6.17: Regression results for two-period ATT estimations across post characteristics including *Economy*, *Health*, *Politics* and *Science*. The estimation models incorporate interaction terms and subsets separately. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

	Interaction	With economy	Without economy		Interaction	With health	Without health
Display	1.172*** (0.025)	1.367*** (0.068)	1.171*** (0.025)	Display	1.231*** (0.025)	0.922*** (0.070)	1.233*** (0.025)
After	0.109*** (0.004)	0.067*** (0.013)	0.111*** (0.004)	After	0.100*** (0.004)	0.031* (0.012)	0.115*** (0.004)
Display×After	−0.959*** (0.005)	−0.856*** (0.014)	−0.959*** (0.005)	Display×After	−0.963*** (0.005)	−0.797*** (0.013)	−0.963*** (0.005)
PostAge	−0.797*** (0.002)	−0.782*** (0.008)	−0.799*** (0.003)	PostAge	−0.798*** (0.002)	−0.672*** (0.007)	−0.813*** (0.003)
Economy	−0.606*** (0.049)			Health	0.514*** (0.055)		
Display×Economy	0.191* (0.071)			Display×Health	−0.299*** (0.079)		
After×Economy	−0.023 (0.010)			After×Health	0.061*** (0.010)		
ATT×Economy	0.101*** (0.014)			ATT×Health	0.162*** (0.014)		
Intercept	1.371*** (0.018)	0.772*** (0.048)	1.370*** (0.018)	Intercept	1.242*** (0.017)	1.877*** (0.049)	1.228*** (0.018)
Post-level RE	✓	✓	✓	Post-level RE	✓	✓	✓
#Observations	654,400	80,416	573,984	#Observations	654,400	63,040	591,360

	Interaction	With politics	With politics		Interaction	With science	Without science
Display	1.315*** (0.027)	0.869*** (0.043)	1.320*** (0.028)	Display	1.122*** (0.025)	1.684*** (0.063)	1.122*** (0.025)
After	0.136*** (0.005)	0.035*** (0.008)	0.135*** (0.005)	After	0.093*** (0.004)	0.180*** (0.012)	0.097*** (0.004)
Display×After	−0.991*** (0.005)	−0.838*** (0.008)	−0.991*** (0.005)	Display×After	−0.949*** (0.005)	−0.951*** (0.013)	−0.949*** (0.005)
PostAge	−0.797*** (0.002)	−0.800*** (0.005)	−0.796*** (0.003)	PostAge	−0.797*** (0.002)	−0.769*** (0.007)	−0.800*** (0.003)
Politics	0.631*** (0.038)			Science	−1.092*** (0.047)		
Display×Politics	−0.437*** (0.053)			Display×Science	0.561*** (0.068)		
After×Politics	−0.106*** (0.007)			After×Science	0.114*** (0.010)		
ATT×Politics	0.153*** (0.010)			ATT×Science	−0.003 (0.014)		
Intercept	1.130*** (0.019)	1.771*** (0.031)	1.123*** (0.020)	Intercept	1.443*** (0.018)	0.371*** (0.044)	1.440*** (0.018)
Post-level RE	✓	✓	✓	Post-level RE	✓	✓	✓
#Observations	654,400	167,040	487,360	#Observations	654,400	89,600	564,800

S6.8 Analysis of Reposting Mechanisms

S6.8.1 Collection of Reposter Information

We further examined how the display of community notes influenced different types of reposters. Due to the rate limits of the X API, collecting comprehensive information on all reposters in our dataset was challenging. Given this, we randomly selected 2000 posts with displayed notes and 2000 posts without displayed notes to analyze the heterogeneity in the treatment effects of community notes among different types of reposters.

For each post, we used the X Pro API's search/all endpoint to gather information about reposters during each hour within a time window of -4 to $+12$ hours relative to the post's publication. Reposts from the pre-display period (-4 h until the note appears) are used to assess pre-trends, while the ATT estimates are identified from reposts occurring after the display of community notes. Due to API rate limits, we sampled reposters randomly within a 5-minute interval for each one-hour time block. Additionally, for each identified reposter, we used the X Pro API's count/all endpoint to retrieve the total number of interactions (reposts) the reposter had with the original poster of the post over the 28 days *preceding* the publication of the post targeted by a community note. The metric thus does not include interactions with the post itself prior to the addition of a note. Based on this data, we construct a dummy variable set to 1 if the reposter had interacted with the creator of the post prior to reposting the specific post that was flagged with a community note.

Moreover, we used the dataset from [17] to assign each reposter a partisan score, which reflects the extent to which a user tends to follow politicians from the left or right of the political spectrum, and a misinformation exposure score, which measures the likelihood that the politicians and public organizations followed by the user disseminate false information. A reposter is classified as right-leaning if their partisan score is greater than 0, and left-leaning if their partisan score is less than 0. Similarly, a reposter is considered to have high exposure to misinformation if their misinformation exposure score exceeds 0.5, and low exposure if the score is 0.5 or below.

We successfully collected 222,068 reposts with reposter information for 3163 source posts. Each reposter has four characteristics: (i) the prior interaction status, indicating whether the reposter had prior interactions with the same author of the misleading post, (ii) the verified status of the reposter, (iii) the political leaning of the reposter, and (iv) the exposure score of the reposter to misinformation. The distributions of the propensity score in the collected posts in the treatment and control groups maintained balanced ($t = 0.113$, $p = 0.910$; see Fig. S6.9).

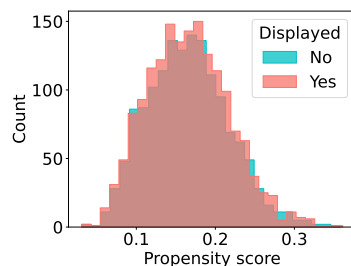


Figure S6.9: The distributions of the propensity score for collected posts in the treatment and control groups.

Table S6.18: Overview of the dataset for reposter analysis. Descriptive statistics are reported for the whole dataset (column (1)), the treated posts that received at least one displayed helpful note (column (2)), and the control posts that did not received a displayed helpful note during our observation period (column (3)).

	All	Posts with displayed notes	
		Yes	No
#Reposts	222,068	106,797	115,271
#Posts	3163	1676	1487
#Posters	1979	1094	1115
<i>Poster characteristics</i>			
<i>LeftLeaning</i>	25.5%	24.6%	26.4%
<i>RightLeaning</i>	30.9%	31.6%	30.2%
<i>HighMisinfoExposure</i>	37.3%	38.2%	36.2%
<i>LowMisinfoExposure</i>	20.6%	19.3%	22.2%
<i>#Reposts across reposter characteristics</i>			
<i>PriorInteraction</i>	122,180	54,697	67,483
<i>Verified</i>	13,461	6253	7208
<i>LeftLeaning</i>	46,916	21,532	25,384
<i>RightLeaning</i>	67,694	32,237	35,457
<i>HighMisinfoExposure</i>	86,867	41,587	45,280
<i>LowMisinfoExposure</i>	135,201	65,210	69,991

S6.8.2 Reposter Dataset

An overview of the dataset for the reposter analysis is shown in Table S6.18. For political orientations, we defined two binary variables, *i.e.*, *LeftLeaning* and *RightLeaning*, to indicate whether the (re)poster is politically left(right)-leaning (= 1) or not (= 0). Moreover, we defined another two binary variables, *i.e.*, *HighMisinfoExposure* and *LowMisinfoExposure*, to indicate whether the (re)poster has high(low) exposure to misinformation (= 1) or not (= 0). Additionally, *Verified* is a binary variable showing that whether the reposter is verified (= 1) or not (= 0). *PriorInteraction* is a binary variable indicating whether the reposter had prior interactions with the authors of misleading posts (= 1) or not (= 0).

S6.8.3 Parallel Trend and Treatment Effect

To validate the robustness of the parallel trend and the treatment effect of community notes, we repeated our multi-period and two-period DiD estimations for the collected post subset. Table S6.19 reports the results for multi-period DiD estimation, showing that the DiD terms during before-display period were not statistically significantly different from zero (each $p > 0.01$). This indicates that the changes of reposts in the treatment group and control group for the post subset remained parallel before the display of community notes. Column (1) of Table S6.20 reports the results for two-period DiD estimation, where the ATT was estimated as -0.633 (99% CI: $[-0.651, -0.614]$; $p < 0.001$). This suggests that the treatment effect of community notes in the subset of posts for reposter analysis is consistent with the overall treatment effect (*i.e.*, -0.612 (99% CI: $[-0.617, -0.608]$; $p < 0.001$). Additionally, we expanded the two-period DiD estimation by incorporating the political leanings and misinformation exposures of posters. As shown in Column (2) of Table S6.20, the ATT estimate remained identical to that in Column (1).

S6.8.4 ATTs Across Reposter Characteristics

Table S6.21 reports the regression results across the four reposter characteristics. The ATT estimate for connected reposters was -0.586 (99% CI: $[-0.608, -0.561]$; $p < 0.001$); the ATT estimate for disconnected reposters was -0.682 (99% CI: $[-0.703, -0.660]$; $p < 0.001$). The ATT estimate for verified reposters was -0.595 (99% CI: $[-0.640, -0.545]$; $p < 0.001$); the ATT estimate for non-verified reposters was -0.634 (99% CI: $[-0.653, -0.615]$; $p < 0.001$). The ATT estimate for left-leaning reposters was -0.625 (99% CI: $[-0.654, -0.595]$; $p < 0.001$); the ATT estimate for right-leaning reposters was -0.626 (99% CI: $[-0.652, -0.598]$; $p < 0.001$). The ATT estimate for high-misinfo-exposure reposters was -0.633 (99% CI: $[-0.655, -0.609]$; $p < 0.001$); the ATT estimate for low-misinfo-exposure reposters was -0.628 (99% CI: $[-0.649, -0.606]$; $p < 0.001$). The efficacy of community notes was 16.4% higher for disconnected reposters compared to connected reposters.

S6.8.5 Sensitivity Across Poster and Reposter Characteristics

Figs. S6.10–S6.13 show the estimated ATTs across different groups of reposters based on posts separated by poster and post characteristics. We found that the trend of higher efficacy of community notes for disconnected reposters than for connected reposters was present across all poster and post characteristics.

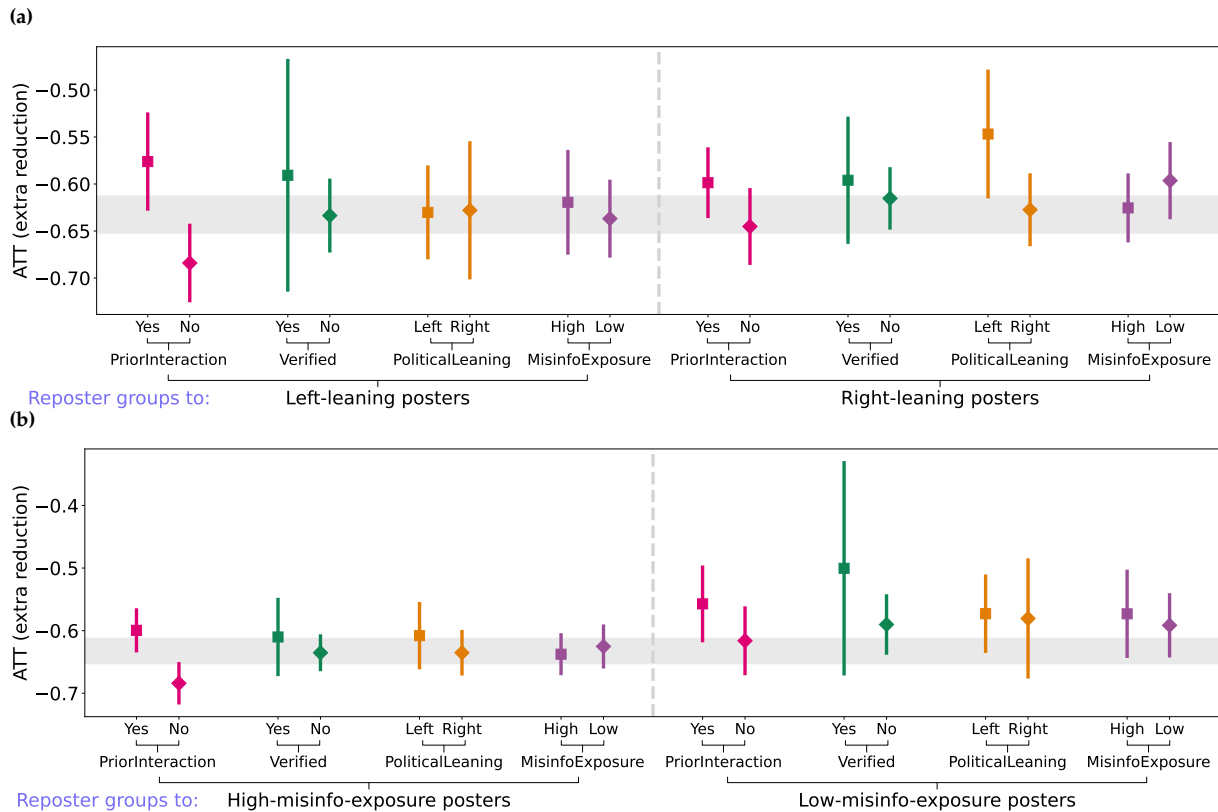


Figure S6.10: The estimated ATTs across different groups of reposters based on posts separated by poster characteristics. (a) Left-leaning vs. right-leaning posters. (b) High-misinfo-exposure vs. low-misinfo-exposure posters. Shown are mean values with 99% CIs. The full estimation results are reported in Tables S6.22–S6.25.

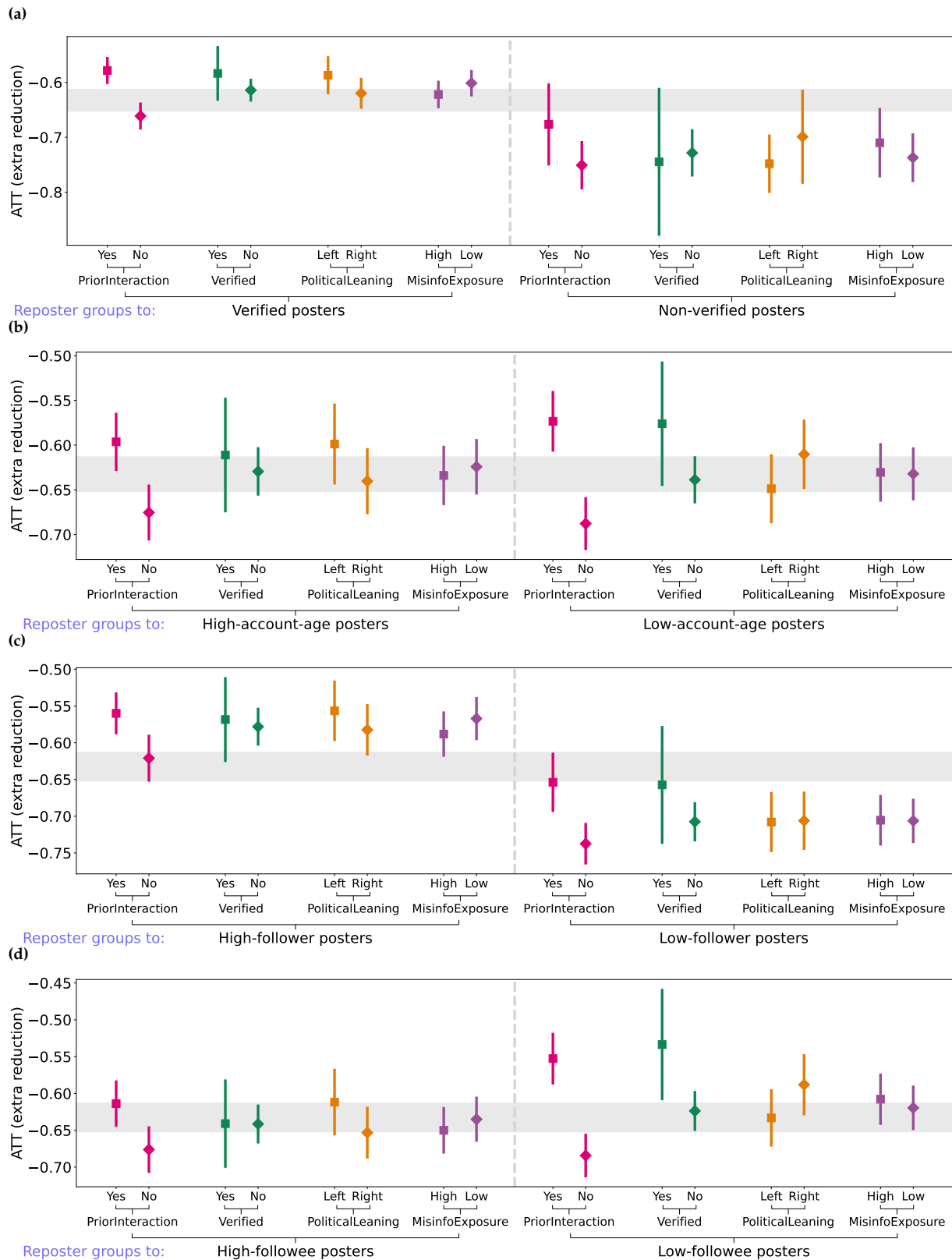


Figure S6.11: The estimated ATTs across different groups of reposter based on posts separated by poster characteristics. (a) Verified vs. non-verified posters. (b) High-account-age vs. low-account-age posters. (c) High-follower vs. low-follower posters. (d) High-follower vs. low-follower posters. Shown are mean values with 99% CIs. The full estimation results are reported in Tables S6.26–S6.33.

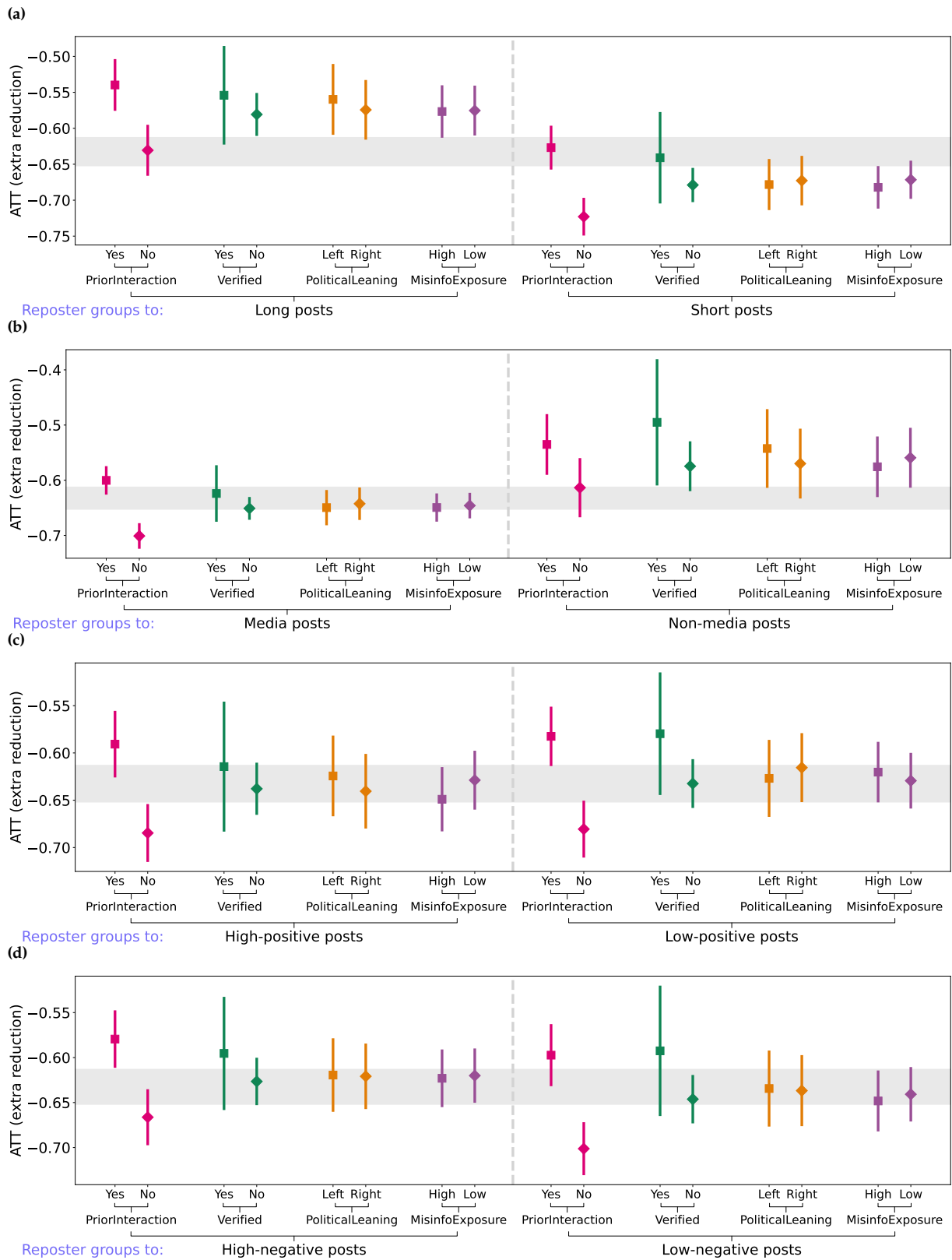


Figure S6.12: The estimated ATTs across different groups of reposters based on posts separated by post characteristics. (a) Long vs. short posts. (b) Media vs. non-media posts. (c) High-positive vs. low-positive posts. (d) High-negative vs. low-negative posts. Shown are mean values with 99% CIs. The full estimation results are reported in Tables S6.34–S6.41.

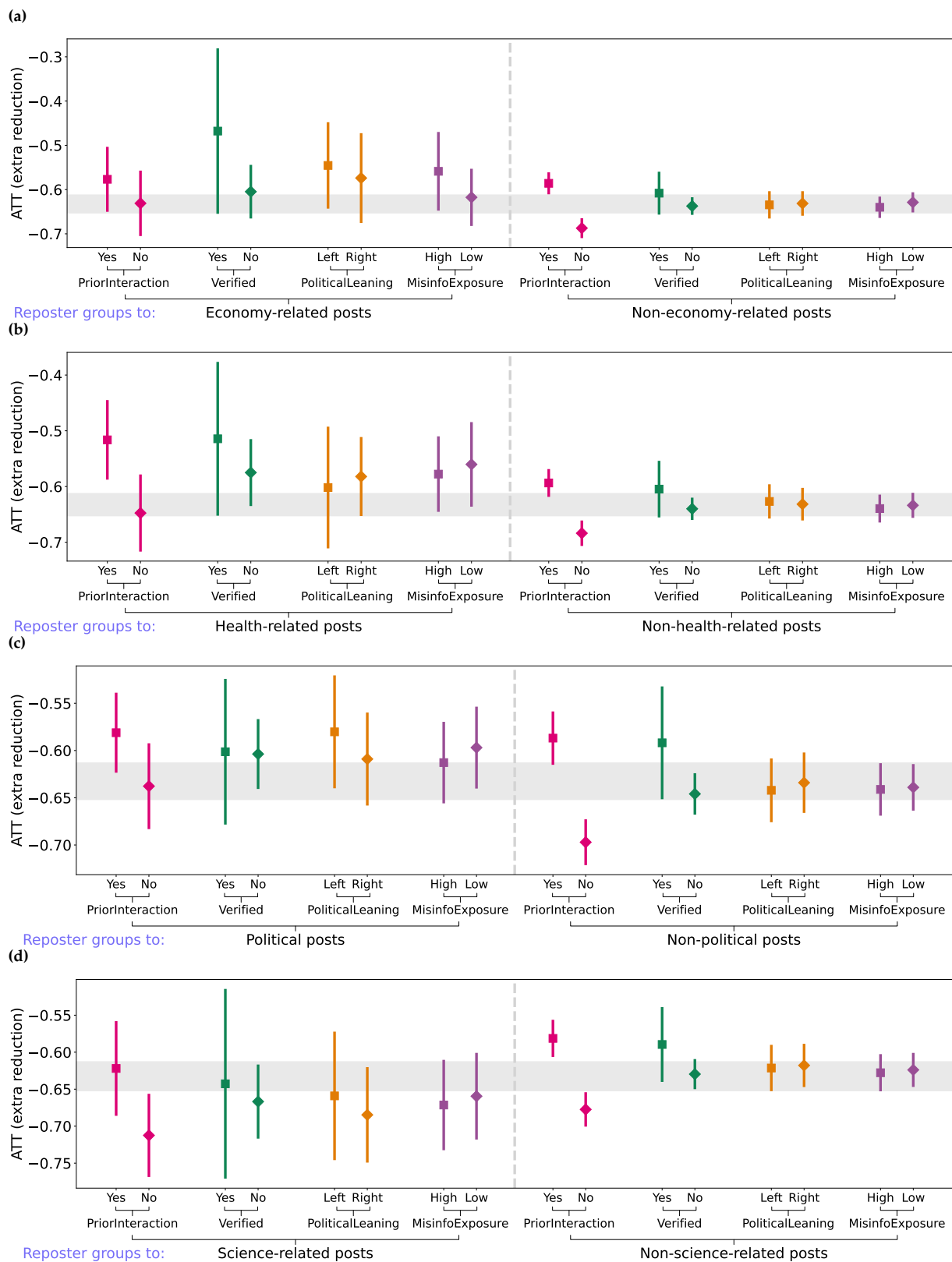


Figure S6.13: The estimated ATTs across different groups of reposters based on posts separated by post characteristics. (a) Economy-related vs. non-economy-related posts. (b) Health-related vs. non-health-related posts. (c) Political vs. non-political posts. (d) Science-related vs. non-science-related posts. Shown are mean values with 99% CIs. The full estimation results are reported in Tables S6.42–S6.49.

Table S6.19: Regression results for parallel test and multi-period ATTs estimation in the subset of posts for reposter analysis. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

Multi-period		Continued	
Display	0.588*** (0.071)	Display×Before:4	0.038 (0.044)
Before:4	0.242*** (0.038)	Display×Before:3	0.020 (0.044)
Before:3	0.145*** (0.036)	Display×Before:2	0.064 (0.044)
Before:2	0.044 (0.034)	Display×After:1	−0.619*** (0.047)
After:1	−0.131*** (0.035)	Display×After:2	−0.879*** (0.048)
After:2	−0.170*** (0.037)	Display×After:3	−0.913*** (0.049)
After:3	−0.227*** (0.040)	Display×After:4	−1.032*** (0.049)
After:4	−0.240*** (0.043)	Display×After:5	−1.045*** (0.049)
After:5	−0.257*** (0.046)	Display×After:6	−1.044*** (0.050)
After:6	−0.329*** (0.050)	Display×After:7	−1.096*** (0.051)
After:7	−0.357*** (0.055)	Display×After:8	−1.029*** (0.051)
After:8	−0.440*** (0.060)	Display×After:9	−1.061*** (0.052)
After:9	−0.472*** (0.064)	Display×After:10	−1.084*** (0.052)
After:10	−0.520*** (0.069)	Display×After:11	−1.039*** (0.053)
After:11	−0.596*** (0.074)	Display×After:12	−1.065*** (0.054)
After:12	−0.617*** (0.079)	PostAge	−0.170*** (0.042)

Continued	
Intercept	0.160* (0.059)
Post-level RE	✓
#Observations	50,608
#Posts	3163

Table S6.20: Regression results for two-period ATT estimation in the subset of posts for reposter analysis. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

	(1) Main	(2) Extended
Display	0.631*** (0.067)	0.621*** (0.065)
After	−0.007 (0.019)	−0.009 (0.019)
Display×After	−1.002*** (0.020)	−1.002*** (0.020)
PostAge	−0.590*** (0.011)	−0.588*** (0.011)
Verified		0.531*** (0.084)
AccountAge		−0.103** (0.036)
Followers		0.268*** (0.041)
Followees		0.079* (0.031)
Words		−0.066 (0.034)
Media		0.435*** (0.075)
Economy		−0.214 (0.099)
Health		0.211 (0.104)
Politics		0.002 (0.077)
Science		−0.339*** (0.101)
Positive		−0.068 (0.041)
Negative		0.133*** (0.040)
MFRO		−0.141*** (0.032)
LeftLeaning _{poster}		0.268 (0.183)
RightLeaning _{poster}		0.397 (0.187)
HighMisinfoExposure _{poster}		−0.124 (0.184)
LowMisinfoExposure _{poster}		−0.324 (0.193)
Intercept	−0.096 (0.050)	−0.885*** (0.111)
Post-level RE	✓	✓
#Observations	50,608	50,608
#Posts	3163	3163

	PriorInteraction		Verified		Political leaning		Misinformation exposure	
	(1) Yes	(2) No	(3) Yes	(4) No	(5) Left	(6) Right	(7) High	(8) Low
Display	0.334*** (0.078)	0.865*** (0.075)	0.463*** (0.082)	0.637*** (0.067)	0.484*** (0.078)	0.742*** (0.079)	0.688*** (0.074)	0.604*** (0.069)
After	-0.043 (0.021)	0.072** (0.025)	-0.111 * (0.042)	-0.001 (0.019)	-0.017 (0.029)	-0.005 (0.027)	-0.005 (0.024)	-0.006 (0.021)
Display × After	-0.881*** (0.022)	-1.145*** (0.026)	-0.904*** (0.045)	-1.006*** (0.020)	-0.981*** (0.030)	-0.984*** (0.028)	-1.001*** (0.025)	-0.989*** (0.022)
PostAge	-0.632*** (0.013)	-0.502*** (0.015)	-0.570*** (0.026)	-0.582*** (0.012)	-0.557*** (0.018)	-0.579*** (0.016)	-0.577*** (0.015)	-0.547*** (0.013)
Intercept	-0.988*** (0.058)	-1.275*** (0.057)	-3.013*** (0.071)	-0.181 *** (0.050)	-1.965*** (0.061)	-1.837*** (0.062)	-1.251*** (0.056)	-0.696*** (0.052)
Post-level RE	✓	✓	✓	✓	✓	✓	✓	✓
#Observations	50,608	50,608	50,608	50,608	50,608	50,608	50,608	50,608

Table S6.21: Regression results for two-period ATT estimations across reposter characteristics. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

	PriorInteraction		Verified		Political leaning		Misinformation exposure	
	(1) Yes	(2) No	(3) Yes	(4) No	(5) Left	(6) Right	(7) High	(8) Low
Display	0.314 (0.145)	0.958*** (0.146)	0.470** (0.165)	0.680*** (0.130)	0.419** (0.144)	1.053*** (0.155)	0.823*** (0.139)	0.624*** (0.133)
After	-0.053 (0.045)	0.092 (0.049)	-0.087 (0.104)	0.009 (0.039)	-0.032 (0.049)	0.037 (0.073)	0.018 (0.054)	-0.001 (0.042)
Display × After	-0.858*** (0.048)	-1.152*** (0.051)	-0.893*** (0.117)	-1.004*** (0.042)	-0.995*** (0.052)	-0.989*** (0.077)	-0.966*** (0.057)	-1.013*** (0.044)
PostAge	-0.695*** (0.029)	-0.555*** (0.030)	-0.583*** (0.064)	-0.620*** (0.024)	-0.627*** (0.031)	-0.567*** (0.045)	-0.582*** (0.034)	-0.606*** (0.026)
Intercept	-1.268*** (0.108)	-1.325*** (0.110)	-3.507*** (0.150)	-0.365*** (0.096)	-1.280*** (0.109)	-2.887*** (0.130)	-1.818*** (0.108)	-0.640*** (0.099)
Post-level RE	✓	✓	✓	✓	✓	✓	✓	✓
#Observations	12,896	12,896	12,896	12,896	12,896	12,896	12,896	12,896

Table S6.22: Regression results for two-period ATT estimations, based on posts from left-leaning posters and across reposter characteristics. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

	PriorInteraction		Verified		Political leaning		Misinformation exposure	
	(1) Yes	(2) No	(3) Yes	(4) No	(5) Left	(6) Right	(7) High	(8) Low
Display	0.089 (0.138)	0.838*** (0.133)	0.234 (0.134)	0.474*** (0.121)	0.438** (0.138)	0.473*** (0.130)	0.490*** (0.125)	0.445*** (0.127)
After	-0.002 (0.034)	0.056 (0.043)	-0.068 (0.061)	0.008 (0.032)	0.024 (0.055)	-0.012 (0.039)	-0.006 (0.036)	0.025 (0.037)
Display × After	-0.913*** (0.036)	-1.036*** (0.045)	-0.906*** (0.065)	-0.955*** (0.033)	-0.791*** (0.059)	-0.987*** (0.040)	-0.982*** (0.038)	-0.907*** (0.039)
PostAge	-0.634*** (0.021)	-0.517*** (0.026)	-0.652*** (0.039)	-0.594*** (0.019)	-0.530*** (0.034)	-0.618*** (0.024)	-0.610*** (0.022)	-0.552*** (0.023)
Intercept	-0.457*** (0.104)	-1.178*** (0.102)	-2.208*** (0.109)	0.110 (0.091)	-2.252*** (0.111)	-0.741*** (0.099)	-0.499*** (0.095)	-0.716 *** (0.096)
Post-level RE	✓	✓	✓	✓	✓	✓	✓	✓
#Observations	15,648	15,648	15,648	15,648	15,648	15,648	15,648	15,648

Table S6.23: Regression results for two-period ATT estimations, based on posts from right-leaning posters and across reposter characteristics. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

	PriorInteraction		Verified		Political leaning		Misinformation exposure	
	(1) Yes	(2) No	(3) Yes	(4) No	(5) Left	(6) Right	(7) High	(8) Low
Display	0.104 (0.125)	0.896*** (0.122)	0.291 (0.126)	0.506*** (0.109)	0.503*** (0.126)	0.431*** (0.122)	0.484*** (0.115)	0.507*** (0.114)
After	-0.050 (0.032)	0.083 (0.040)	-0.080 (0.058)	-0.011 (0.030)	-0.027 (0.051)	-0.026 (0.037)	-0.017 (0.034)	-0.006 (0.035)
Display × After	-0.915*** (0.034)	-1.152*** (0.042)	-0.942*** (0.062)	-1.009*** (0.031)	-0.937*** (0.053)	-1.009*** (0.039)	-1.015*** (0.036)	-0.982*** (0.036)
PostAge	-0.623*** (0.020)	-0.524*** (0.024)	-0.633*** (0.037)	-0.588*** (0.018)	-0.549*** (0.031)	-0.597*** (0.023)	-0.593*** (0.021)	-0.552*** (0.021)
Intercept	-0.536*** (0.094)	-1.229*** (0.094)	-2.397*** (0.104)	0.060 (0.082)	-2.132*** (0.101)	-0.949*** (0.093)	-0.638*** (0.088)	-0.686*** (0.087)
Post-level RE	✓	✓	✓	✓	✓	✓	✓	✓
#Observations	18,864	18,864	18,864	18,864	18,864	18,864	18,864	18,864

Table S6.24: Regression results for two-period ATT estimations, based on posts from high-misinformation-exposure posters and across reposter characteristics. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

Table S6.25: Regression results for two-period ATT estimations, based on posts from low-misinformation-exposure posters and across reposter characteristics. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

[illegible]

Table S6.26: Regression results for two-period ATT estimations, based on posts from verified posters and across reposter characteristics. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

[illegible]

Table S6.27: Regression results for two-period ATT estimations, based on posts from non-verified posters and across reporter characteristics. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

[illegible]

Table S6.28: Regression results for two-period ATT estimations, based on posts from older accounts and across reposter characteristics. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

[illegible]

	PriorInteraction		Verified		Political leaning		Misinformation exposure	
	(1) Yes	(2) No	(3) Yes	(4) No	(5) Left	(6) Right	(7) High	(8) Low
Display	0.293* (0.114)	0.992*** (0.106)	0.464*** (0.118)	0.685*** (0.098)	0.655*** (0.111)	0.664*** (0.114)	0.668*** (0.106)	0.681*** (0.100)
After	-0.064 (0.029)	0.059 (0.035)	-0.109 (0.059)	-0.017 (0.027)	0.000 (0.040)	-0.030 (0.037)	-0.025 (0.033)	-0.012 (0.030)
Display × After	-0.852*** (0.031)	-1.164*** (0.037)	-0.858*** (0.064)	-1.018*** (0.028)	-1.046*** (0.043)	-0.942*** (0.039)	-0.996*** (0.034)	-1.000*** (0.031)
PostAge	-0.600*** (0.018)	-0.481*** (0.021)	-0.516*** (0.037)	-0.563*** (0.016)	-0.556*** (0.025)	-0.528*** (0.022)	-0.538*** (0.020)	-0.538*** (0.018)
Intercept	-0.949*** (0.084)	-1.199*** (0.080)	-3.005*** (0.101)	-0.122 (0.072)	-2.068*** (0.087)	-1.703*** (0.087)	-1.173*** (0.080)	-0.646*** (0.074)
Post-level RE	✓	✓	✓	✓	✓	✓	✓	✓
#Observations	25,312	25,312	25,312	25,312	25,312	25,312	25,312	25,312

Table S6.29: Regression results for two-period ATT estimations, based on posts from younger accounts and across reposter characteristics. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

	PriorInteraction		Verified		Political leaning		Misinformation exposure	
	(1) Yes	(2) No	(3) Yes	(4) No	(5) Left	(6) Right	(7) High	(8) Low
Display	0.218 (0.096)	0.504*** (0.108)	0.243 (0.102)	0.328*** (0.093)	0.120 (0.104)	0.474*** (0.106)	0.409*** (0.098)	0.280** (0.096)
After	-0.047 (0.024)	0.077 (0.031)	-0.057 (0.048)	-0.012 (0.023)	-0.024 (0.034)	-0.010 (0.031)	-0.006 (0.028)	-0.020 (0.025)
Display × After	-0.821*** (0.025)	-0.970*** (0.033)	-0.840*** (0.052)	-0.863*** (0.024)	-0.813*** (0.036)	-0.873*** (0.033)	-0.887*** (0.029)	-0.837*** (0.026)
PostAge	-0.635*** (0.015)	-0.507*** (0.019)	-0.662*** (0.031)	-0.598*** (0.014)	-0.555*** (0.021)	-0.634*** (0.019)	-0.620*** (0.017)	-0.553*** (0.015)
Intercept	0.044 (0.072)	-1.046*** (0.082)	-2.280*** (0.084)	0.425*** (0.070)	-1.361*** (0.080)	-1.068*** (0.081)	-0.541*** (0.074)	-0.131 (0.072)
Post-level RE	✓	✓	✓	✓	✓	✓	✓	✓
#Observations	25,296	25,296	25,296	25,296	25,296	25,296	25,296	25,296

Table S6.30: Regression results for two-period ATT estimations, based on posts from accounts with higher follower counts and across reposter characteristics. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

	PriorInteraction		Verified		Political leaning		Misinformation exposure	
	(1) Yes	(2) No	(3) Yes	(4) No	(5) Left	(6) Right	(7) High	(8) Low
Display	0.347** (0.108)	1.209*** (0.104)	0.644*** (0.130)	0.917*** (0.092)	0.855*** (0.115)	0.977*** (0.113)	0.933*** (0.104)	0.913*** (0.096)
After	-0.066 (0.041)	0.061 (0.040)	-0.269** (0.083)	0.002 (0.033)	-0.031 (0.051)	-0.006 (0.050)	-0.024 (0.043)	0.005 (0.038)
Display × After	-1.061*** (0.045)	-1.337*** (0.042)	-1.071*** (0.091)	-1.229*** (0.035)	-1.231*** (0.054)	-1.225*** (0.052)	-1.222*** (0.045)	-1.225*** (0.040)
PostAge	-0.591*** (0.027)	-0.497*** (0.024)	-0.320*** (0.048)	-0.544*** (0.020)	-0.544*** (0.032)	-0.447*** (0.030)	-0.476*** (0.026)	-0.527*** (0.023)
Intercept	-1.936*** (0.084)	-1.484*** (0.080)	-3.636*** (0.119)	-0.734*** (0.069)	-2.541*** (0.093)	-2.517*** (0.091)	-1.888*** (0.081)	-1.217*** (0.073)
Post-level RE	✓	✓	✓	✓	✓	✓	✓	✓
#Observations	25,312	25,312	25,312	25,312	25,312	25,312	25,312	25,312

Table S6.31: Regression results for two-period ATT estimations, based on posts from accounts with lower follower counts and across reposter characteristics. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

	PriorInteraction		Verified		Political leaning		Misinformation exposure	
	(1) Yes	(2) No	(3) Yes	(4) No	(5) Left	(6) Right	(7) High	(8) Low
Display	0.380*** (0.103)	0.792*** (0.106)	0.489*** (0.112)	0.632*** (0.092)	0.397*** (0.110)	0.804*** (0.111)	0.744*** (0.101)	0.584*** (0.095)
After	-0.047 (0.030)	-0.002 (0.037)	-0.112 (0.060)	-0.038 (0.028)	-0.081 (0.043)	-0.046 (0.038)	-0.050 (0.034)	-0.032 (0.031)
Display × After	-0.951*** (0.032)	-1.128*** (0.038)	-1.024*** (0.065)	-1.026*** (0.029)	-0.946*** (0.045)	-1.059*** (0.039)	-1.050*** (0.035)	-1.008*** (0.032)
PostAge	-0.643*** (0.019)	-0.498*** (0.022)	-0.585*** (0.037)	-0.588*** (0.017)	-0.555*** (0.027)	-0.578*** (0.023)	-0.581*** (0.021)	-0.547*** (0.019)
Intercept	-0.871*** (0.078)	-1.215*** (0.082)	-2.859*** (0.097)	-0.144 (0.070)	-1.921*** (0.088)	-1.716*** (0.087)	-1.149*** (0.079)	-0.709*** (0.073)
Post-level RE	✓	✓	✓	✓	✓	✓	✓	✓
#Observations	25,296	25,296	25,296	25,296	25,296	25,296	25,296	25,296

Table S6.32: Regression results for two-period ATT estimations, based on posts from accounts with higher followee counts and across reposter characteristics. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

Table S6.33: Regression results for two-period ATT estimations, based on posts from accounts with lower followee counts and across reposter characteristics. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

[illegible]

Table S6.34: Regression results for two-period ATT estimations, based on longer posts and across reposter characteristics. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

[illegible]

Table S6.35: Regression results for two-period ATT estimations, based on shorter posts and across reposter characteristics. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

[illegible]

Table S6.36: Regression results for two-period ATT estimations, based on media posts and across reposter characteristics. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

[illegible]

	PriorInteraction		Verified		Political leaning		Misinformation exposure	
	(1) Yes	(2) No	(3) Yes	(4) No	(5) Left	(6) Right	(7) High	(8) Low
Display	0.184 (0.149)	0.591*** (0.148)	0.250 (0.164)	0.439*** (0.130)	0.198 (0.155)	0.673*** (0.157)	0.544*** (0.142)	0.347 (0.137)
After	-0.042 (0.043)	0.025 (0.051)	-0.210* (0.079)	-0.013 (0.039)	-0.100 (0.056)	-0.026 (0.054)	-0.030 (0.047)	-0.033 (0.045)
Display × After	-0.766*** (0.046)	-0.951*** (0.054)	-0.683*** (0.088)	-0.855*** (0.041)	-0.782*** (0.060)	-0.844*** (0.057)	-0.857*** (0.050)	-0.819*** (0.048)
PostAge	-0.692*** (0.027)	-0.536*** (0.031)	-0.563*** (0.051)	-0.623*** (0.024)	-0.564*** (0.036)	-0.642*** (0.034)	-0.616*** (0.030)	-0.579*** (0.028)
Intercept	-1.238*** (0.110)	-1.462*** (0.112)	-3.105*** (0.139)	-0.399*** (0.096)	-2.051*** (0.119)	-2.131*** (0.121)	-1.408*** (0.107)	-0.984*** (0.102)
Post-level RE	✓	✓	✓	✓	✓	✓	✓	✓
#Observations	13,728	13,728	13,728	13,728	13,728	13,728	13,728	13,728

Table S6.37: Regression results for two-period ATT estimations, based on non-media posts and across reposter characteristics. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

	PriorInteraction		Verified		Political leaning		Misinformation exposure	
	(1) Yes	(2) No	(3) Yes	(4) No	(5) Left	(6) Right	(7) High	(8) Low
Display	0.493*** (0.110)	1.007*** (0.105)	0.689*** (0.121)	0.822*** (0.095)	0.674*** (0.112)	0.904*** (0.114)	0.857*** (0.105)	0.804*** (0.097)
After	-0.057 (0.032)	0.068 (0.036)	-0.132 (0.065)	-0.008 (0.028)	-0.013 (0.042)	-0.025 (0.041)	-0.009 (0.036)	-0.003 (0.031)
Display × After	-0.893*** (0.033)	-1.154*** (0.038)	-0.953*** (0.069)	-1.016*** (0.030)	-0.979*** (0.044)	-1.023*** (0.043)	-1.047*** (0.038)	-0.991*** (0.033)
PostAge	-0.569*** (0.020)	-0.461*** (0.021)	-0.515*** (0.039)	-0.531*** (0.017)	-0.537*** (0.026)	-0.503*** (0.025)	-0.511*** (0.022)	-0.511*** (0.019)
Intercept	-1.240*** (0.083)	-1.368*** (0.080)	-3.314*** (0.106)	-0.373*** (0.071)	-2.167*** (0.088)	-2.127*** (0.089)	-1.518*** (0.080)	-0.851*** (0.073)
Post-level RE	✓	✓	✓	✓	✓	✓	✓	✓
#Observations	25,296	25,296	25,296	25,296	25,296	25,296	25,296	25,296

Table S6.38: Regression results for two-period ATT estimations, based on high-positive posts and across reposter characteristics. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

	PriorInteraction		Verified		Political leaning		Misinformation exposure	
	(1) Yes	(2) No	(3) Yes	(4) No	(5) Left	(6) Right	(7) High	(8) Low
Display	0.172 (0.109)	0.722*** (0.107)	0.254 (0.112)	0.448*** (0.096)	0.297* (0.109)	0.581*** (0.110)	0.519*** (0.103)	0.402*** (0.099)
After	-0.028 (0.027)	0.079 (0.035)	-0.093 (0.055)	0.010 (0.026)	-0.019 (0.040)	0.014 (0.035)	0.002 (0.031)	-0.004 (0.029)
Display × After	-0.873*** (0.029)	-1.141*** (0.037)	-0.867*** (0.060)	-1.001*** (0.027)	-0.986*** (0.042)	-0.956*** (0.037)	-0.968*** (0.033)	-0.993*** (0.031)
PostAge	-0.685*** (0.017)	-0.543*** (0.021)	-0.614*** (0.035)	-0.628*** (0.016)	-0.577*** (0.025)	-0.640*** (0.022)	-0.630*** (0.019)	-0.582*** (0.018)
Intercept	-0.728*** (0.082)	-1.182*** (0.082)	-2.718*** (0.094)	0.017 (0.072)	-1.763*** (0.085)	-1.538*** (0.085)	-0.978*** (0.078)	-0.539*** (0.074)
Post-level RE	✓	✓	✓	✓	✓	✓	✓	✓
#Observations	25,312	25,312	25,312	25,312	25,312	25,312	25,312	25,312

Table S6.39: Regression results for two-period ATT estimations, based on low-positive posts and across reposter characteristics. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

	PriorInteraction		Verified		Political leaning		Misinformation exposure	
	(1) Yes	(2) No	(3) Yes	(4) No	(5) Left	(6) Right	(7) High	(8) Low
Display	0.098 (0.111)	0.624*** (0.106)	0.238 (0.112)	0.400*** (0.096)	0.280 (0.110)	0.506*** (0.110)	0.450*** (0.102)	0.370*** (0.100)
After	-0.060 (0.027)	0.014 (0.034)	-0.054 (0.054)	-0.036 (0.026)	-0.059 (0.039)	-0.038 (0.035)	-0.042 (0.031)	-0.038 (0.029)
Display × After	-0.866*** (0.029)	-1.098*** (0.036)	-0.905*** (0.060)	-0.985*** (0.027)	-0.966*** (0.042)	-0.970*** (0.037)	-0.976*** (0.033)	-0.968*** (0.031)
PostAge	-0.657*** (0.017)	-0.502*** (0.021)	-0.612*** (0.035)	-0.593*** (0.016)	-0.559*** (0.025)	-0.586*** (0.022)	-0.584*** (0.019)	-0.562*** (0.018)
Intercept	-0.692*** (0.082)	-1.045*** (0.080)	-2.726*** (0.093)	0.091 (0.071)	-1.699*** (0.085)	-1.470*** (0.084)	-0.899*** (0.077)	-0.478*** (0.075)
Post-level RE	✓	✓	✓	✓	✓	✓	✓	✓
#Observations	25,296	25,296	25,296	25,296	25,296	25,296	25,296	25,296

Table S6.40: Regression results for two-period ATT estimations, based on high-negative posts and across reposter characteristics. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

	PriorInteraction		Verified		Political leaning		Misinformation exposure	
	(1) Yes	(2) No	(3) Yes	(4) No	(5) Left	(6) Right	(7) High	(8) Low
Display	0.392*** (0.083)	0.911*** (0.079)	0.565*** (0.088)	0.690*** (0.072)	0.522*** (0.082)	0.835*** (0.084)	0.771*** (0.078)	0.633*** (0.074)
After	-0.042 (0.023)	0.059 (0.027)	-0.113 (0.047)	-0.004 (0.021)	-0.017 (0.030)	-0.023 (0.030)	-0.014 (0.026)	-0.003 (0.023)
Display × After	-0.900*** (0.024)	-1.151*** (0.028)	-0.928*** (0.050)	-1.022*** (0.022)	-0.985*** (0.032)	-0.999*** (0.031)	-1.020*** (0.027)	-1.004*** (0.024)
PostAge	-0.638*** (0.014)	-0.503*** (0.016)	-0.579*** (0.029)	-0.586*** (0.012)	-0.557*** (0.019)	-0.575*** (0.018)	-0.572*** (0.016)	-0.557*** (0.014)
Intercept	-1.066*** (0.063)	-1.288*** (0.061)	-3.146*** (0.077)	-0.228*** (0.054)	-1.933*** (0.065)	-1.979*** (0.066)	-1.361*** (0.060)	-0.696*** (0.056)
Post-level RE	✓	✓	✓	✓	✓	✓	✓	✓
#Observations	44,992	44,992	44,992	44,992	44,992	44,992	44,992	44,992

Table S6.45: Regression results for two-period ATT estimations, based on non-health-related posts and across reposter characteristics. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

	PriorInteraction		Verified		Political leaning		Misinformation exposure	
	(1) Yes	(2) No	(3) Yes	(4) No	(5) Left	(6) Right	(7) High	(8) Low
Display	0.334 (0.134)	0.773*** (0.139)	0.327 (0.145)	0.572*** (0.123)	0.380 (0.151)	0.735*** (0.147)	0.651*** (0.131)	0.556*** (0.129)
After	-0.078 (0.037)	-0.012 (0.048)	-0.108 (0.070)	-0.059 (0.035)	-0.051 (0.053)	-0.085 (0.047)	-0.049 (0.042)	-0.061 (0.041)
Display × After	-0.870*** (0.039)	-1.015*** (0.049)	-0.920*** (0.075)	-0.926*** (0.036)	-0.868*** (0.055)	-0.939*** (0.049)	-0.949*** (0.043)	-0.909*** (0.042)
PostAge	-0.659*** (0.023)	-0.536*** (0.028)	-0.636*** (0.045)	-0.615*** (0.021)	-0.625*** (0.033)	-0.601*** (0.029)	-0.616*** (0.026)	-0.585*** (0.025)
Intercept	-0.587*** (0.102)	-1.363*** (0.108)	-2.574*** (0.122)	-0.100 (0.093)	-2.026*** (0.119)	-1.500*** (0.114)	-0.931*** (0.101)	-0.800*** (0.098)
Post-level RE	✓	✓	✓	✓	✓	✓	✓	✓
#Observations	14,288	14,288	14,288	14,288	14,288	14,288	14,288	14,288

Table S6.46: Regression results for two-period ATT estimations, based on political posts and across reposter characteristics. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

	PriorInteraction		Verified		Political leaning		Misinformation exposure	
	(1) Yes	(2) No	(3) Yes	(4) No	(5) Left	(6) Right	(7) High	(8) Low
Display	0.322*** (0.094)	0.902*** (0.089)	0.512*** (0.099)	0.661*** (0.081)	0.525*** (0.091)	0.737*** (0.094)	0.696*** (0.088)	0.624*** (0.082)
After	-0.028 (0.025)	0.103*** (0.030)	-0.116 (0.052)	0.023 (0.023)	-0.004 (0.034)	0.033 (0.032)	0.014 (0.029)	0.015 (0.025)
Display × After	-0.884*** (0.026)	-1.194*** (0.031)	-0.896*** (0.057)	-1.038*** (0.024)	-1.028*** (0.037)	-1.005*** (0.034)	-1.025*** (0.030)	-1.019*** (0.026)
PostAge	-0.617*** (0.016)	-0.492*** (0.017)	-0.529*** (0.032)	-0.568*** (0.014)	-0.530*** (0.021)	-0.565*** (0.020)	-0.557*** (0.018)	-0.535*** (0.015)
Intercept	-1.154*** (0.070)	-1.243*** (0.068)	-3.183*** (0.087)	-0.214*** (0.060)	-1.942*** (0.071)	-1.968*** (0.073)	-1.379*** (0.067)	-0.658*** (0.061)
Post-level RE	✓	✓	✓	✓	✓	✓	✓	✓
#Observations	36,320	36,320	36,320	36,320	36,320	36,320	36,320	36,320

Table S6.47: Regression results for two-period ATT estimations, based on non-political posts and across reposter characteristics. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

	PriorInteraction		Verified		Political leaning		Misinformation exposure	
	(1) Yes	(2) No	(3) Yes	(4) No	(5) Left	(6) Right	(7) High	(8) Low
Display	0.424 (0.211)	0.899*** (0.207)	0.407 (0.218)	0.739*** (0.186)	0.619** (0.217)	0.791*** (0.213)	0.795*** (0.204)	0.698*** (0.188)
After	0.085 (0.063)	0.228** (0.072)	-0.182 (0.126)	0.196*** (0.055)	0.038 (0.092)	0.230** (0.076)	0.182* (0.069)	0.154 (0.063)
Display × After	-0.973*** (0.066)	-1.246*** (0.076)	-1.029*** (0.139)	-1.099*** (0.058)	-1.076*** (0.099)	-1.154*** (0.079)	-1.113*** (0.072)	-1.077*** (0.067)
PostAge	-0.639*** (0.040)	-0.516*** (0.043)	-0.473*** (0.079)	0.607*** (0.034)	0.532*** (0.058)	-0.587*** (0.047)	-0.592*** (0.043)	-0.563*** (0.039)
Intercept	-1.616*** (0.157)	-1.638*** (0.155)	-3.020*** (0.186)	-0.740*** (0.136)	-2.429*** (0.171)	-2.190*** (0.164)	-1.781*** (0.154)	-1.183*** (0.140)
Post-level RE	✓	✓	✓	✓	✓	✓	✓	✓
#Observations	6480	6480	6480	6480	6480	6480	6480	6480

Table S6.48: Regression results for two-period ATT estimations, based on science-related posts and across reposter characteristics. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

Table S6.49: Regression results for two-period ATT estimations, based on non-science-related posts and across reposter characteristics. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

[illegible]

S6.9 Robustness Checks

S6.9.1 ATT Estimation With Post-Level Fixed Effects

Table S6.50 reports regression results for two-period ATT estimations based on the dataset for main analysis (Column (1)) and the dataset for reposter analysis (Column (2)), respectively. Using data for main analysis, the ATT estimate was -0.567 (99% CI: $[-0.571, -0.563]$; $p < 0.001$), slightly lower than the ATT estimate with post-level random effects (ATT: -0.612 ; 99% CI: $[-0.617, -0.608]$; $p < 0.001$). However, the ATT estimation with post-level fixed effects dropped 25,200 repost count observations across 1575 posts due to all zero outcomes in the dataset for main analysis. Based on the dataset for reposter analysis where 464 repost observations across 29 posts were dropped due to all zero outcomes during fixed-effects model estimation, the ATT estimate was -0.613 (99% CI: $[-0.629, -0.595]$; $p < 0.001$), which was similar to the ATT estimates with post-level random effects based on the datasets for both main analysis and reposter analysis.

S6.9.2 ATT Estimation With Zero-Inflated Negative Binomial Regression

We used zero-inflated negative binomial regression models to account for potential excess zeros with the increase of post age. Table S6.50 reports regression results for two-period ATT estimations based on the dataset for main analysis (Column (3)) and the dataset for reposter analysis (Column (4)), respectively. Based on the posts for main analysis, the ATT estimate was -0.584 (99% CI: $[-0.604, -0.562]$; $p < 0.001$), closely aligning with the main ATT estimate in the mixed negative binomial regression model (ATT: -0.612 ; 99% CI: $[-0.617, -0.608]$; $p < 0.001$). Based on the posts for reposter analysis, the ATT estimate was -0.556 (99% CI: $[-0.622, -0.480]$; $p < 0.001$), consistent with the ATT estimates in the mixed negative binomial regression models based on the datasets for both main analysis and reposter analysis.

In summary, our results are consistent and robust across different modeling approaches and datasets.

S6.9.3 Analysis With Restrictions of Note Scores and Pre-Display Engagement

To further ensure the robustness of our estimates, we conducted an additional robustness check restricting the sample based on note scores and pre-display engagement. Specifically, we applied two restrictions. First, we restricted our analysis to posts with Community Notes that fall within a narrow window around the display threshold. This restriction can help mitigate unobserved confounding factors related to note scores, though it significantly reduces the number of posts during estimation. For our analysis, we selected the window of note helpfulness scores between 0.35 and 0.45, where 0.40 served as the display threshold. Within this selection window, we found 5352 posts in the treatment group, while only 971 posts were in the control group. Second, to balance the treatment and control groups and ensure that variations in post engagement following the treatment are not influenced by pre-existing disparities prior to the intervention, we matched each post within the control group with a

Table S6.50: Regression results for two-period ATT estimation with post-level fixed effects based on the dataset for main analysis (Column (1)) and the subset for reposter analysis (Column (2)). Regression results for two-period ATT estimation with zero-inflated negative binomial regressions based on the dataset for main analysis (Column (3)) and the subset for reposter analysis (Column (4)). Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

	Post-level fixed effects		Zero-inflated negative binomial	
	(1) Data for main analysis	(2) Data for reposter analysis	(3) Data for main analysis	(4) Data for reposter analysis
Display	0.303*** (0.006)	0.191*** (0.035)	0.373*** (0.028)	0.249*** (0.072)
After	−0.158*** (0.003)	−0.101*** (0.016)	−0.517*** (0.014)	−0.465*** (0.035)
Display×After	−0.836*** (0.004)	−0.948*** (0.017)	−0.876*** (0.020)	−0.813*** (0.062)
PostAge	−0.401*** (0.002)	−0.413*** (0.010)		
Intercept	1.106*** (0.005)	1.348*** (0.027)	4.291*** (0.023)	1.949*** (0.058)
Inflate				
PostAge			2.121*** (0.023)	1.600*** (0.091)
Intercept			−5.207*** (0.059)	−3.961*** (0.226)
Post-level FE	✓	✓		
#Observations	629,200	50,144	654,400	50,608
#Posts	39,325	3134	40,900	3163

corresponding post from the treatment group based on the pre-display repost activity. We retained matched pairs where the number of reposts during the pre-display period did not differ by more than 0.1%, which resulted in 813 pairs.

Within this restricted sample, we assessed the observed diffusion of treated posts in comparison to the counterfactual diffusion of their matched counterparts. Specifically, we estimated the treatment effect using two complementary approaches: a Difference-in-Differences (DiD) design, analogous to our main analysis, and a Regression Discontinuity Design (RDD) based on note helpfulness scores.

DiD estimation: The DiD estimation results based on the matched pairs are reported in Column (1) of Table S6.51. Fig. S6.14a shows the estimates for the multi-period and two-period ATTs. The two-period ATT estimate was −0.595 (99% CI: [−0.617, −0.572]; $p < 0.001$), which is very close to the two-period ATT estimate in the main analysis, reinforcing the robustness of our findings.

Regression Discontinuity Design (RDD): As an additional check, we used the note scores as the running variable to estimate the treatment effect of community notes based on a Regression Discontinuity Design (RDD). Specifically, we evaluated the changes in repost count within the paired posts over the note helpfulness scores between and after display. As shown in Fig. S6.14b, the data shows a clear drop in after-display reposts (the total number of reposts during the 12-hour after-display period) after the cutoff point of 0.40, while the before-display reposts (the total number of reposts during the 4-hour before-display period) remained relatively stable.

We specified the following RDD model to estimate the reduction in

reposts after the cutoff relative to before:

$$\begin{aligned} \log(E(\text{RepostCount}|x_{it})) = & \beta_0 + \beta_1 \text{After}_t + \beta_2 \text{Helpful}_i \\ & + \beta_3 \text{After}_t \times \text{Helpful}_i + \beta_4 \text{NoteScore}_i \quad (\text{S6.2}) \\ & + \beta_5 \text{PostAge}_{it} + \mu_{post}, \end{aligned}$$

where After_t is a binary variable indicating whether the reposts were created after the display of community notes ($= 1$) or not ($= 0$). Additionally, the binary variable Helpful_i indicates whether the notes associated with the source posts were eligible to earn helpful status based on the cutoff point. Therefore, the coefficient estimate for the interaction term $\text{After}_t \times \text{Helpful}_i$, *i.e.*, β_3 represents the estimated treatment effect of community notes using RDD.

Note that the helpfulness scores in our dataset were computed from publicly available log data [299] and may differ slightly from those in the production environment, as ratings are updated dynamically. Consequently, a small number of posts with scores below 0.40 appear in the treatment group, and some with scores above 0.40 appear in the control group. We excluded these posts (188 below-threshold treatment posts and 247 above-threshold control posts) before estimating the model.

The estimation results are reported in Column (2) of Table S6.51. The RDD estimate for the treatment effect was -0.616 (99% CI: $[-0.640, -0.591]$; $p < 0.001$), which is nearly identical to the estimated effect using DiD in the main analysis.

Together, these findings provide strong confirmatory evidence that our estimates are not driven by systematic differences in post characteristics. Instead, the observed decline in engagement is attributable to the presence of Community Notes rather than underlying differences between noted and non-noted posts.

S6.9.4 Comparison to Previous Work Analyzing Aggregated Repost Counts

A previous study examined cumulative repost counts (*i.e.*, aggregated engagement data rather than time-series data) for source posts created from the roll-out of Community Notes program until April 2023 and found no evidence showing that community notes reduced overall engagement with misleading posts [74]. One possible explanation is that the display of community notes might have been too slow to during the early stages of the program to produce a measurable effect on overall engagement. In contrast, our analysis – based on a longer observation timeframe until June 2024 – implies that community notes reduced the overall number of reposts for misleading posts on X by 14.9%.

To reconcile these findings and assess consistency between the two studies, we repeated our analysis based on posts created at the same timeframe as in [74]. The DiD estimation results are reported in Column (3) of Table S6.51. The two-period ATT estimate was -0.491 (99% CI: $[-0.510, -0.471]$; $p < 0.001$), indicating a weaker treatment effect during the early stages of the program. This result is consistent with our previous sensitivity analyses (see Suppl. S6.7), which showed that the treatment effect of community notes increased over time.

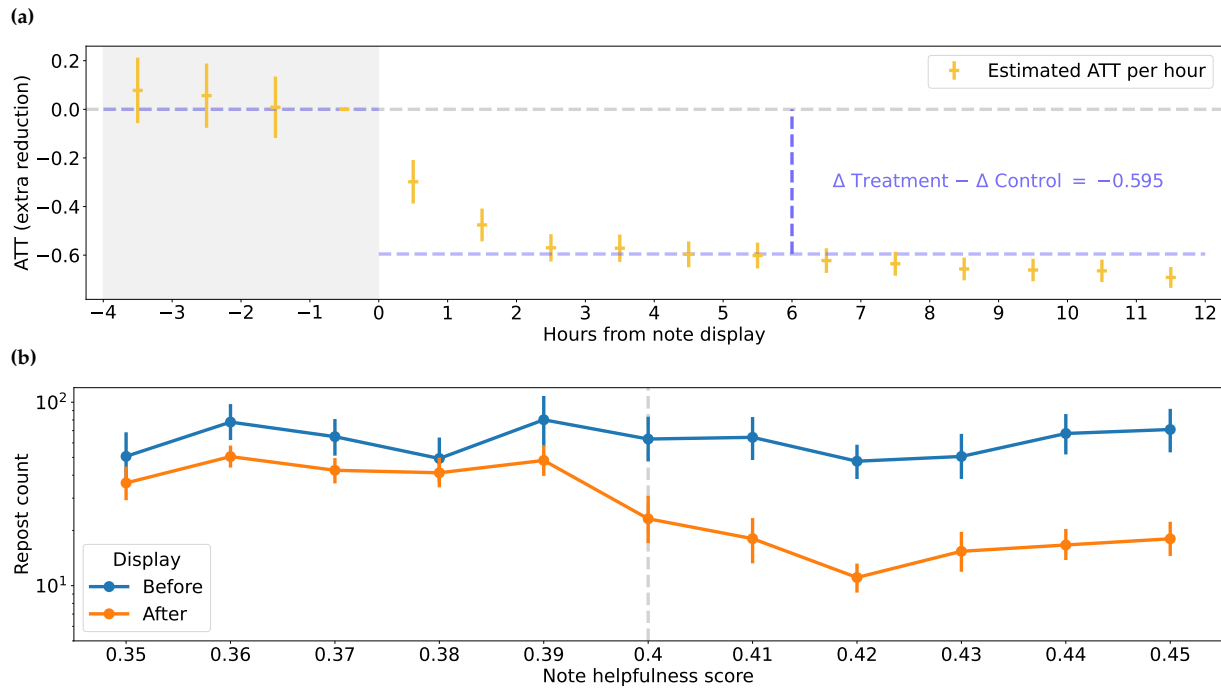


Figure S6.14: Analysis with restrictions of note scores and pre-display engagement. (a) Two-period (purple) and multi-period (yellow) ATTs estimated using DiD design. (b) Changes in repost counts before and after the display of Community Notes across note helpfulness scores. The before-display reposts represent the total number of reposts during the 4-hour before-display period, and the after-display reposts represent the total number of reposts during the 12-hour after-display period. The error bars represent 99% CIs.

Next, we compared the ratio of the predicted cumulative repost count in the absence of displayed notes to the actual cumulative number of reposts. For posts created up to April 2023 (the period analyzed in [74]), the average reduction in reposts was only 7.2%, with half of the posts showing no more than a 1.5% decrease. This is substantially smaller than the reduction in total reposts observed during our full study period (mean of 14.9%; median of 10.1%). This can primarily be explained by two factors: (i) the post-level effect was smaller at this stage (49.1% vs. 61.2%), and (ii) the time-to-display was substantially longer (median of 27.8 hours vs. 18.1 hours), reducing the system-wide effect. The particularly small median reduction during the early stages of Community Notes program further indicates that the observed reduction was primarily driven by a few posts, rather than a broad reduction across all posts.

According to a two-sided KS-test, during the timeframe analyzed in [74], the observed repost count for source posts with displayed community notes (mean = 2792) was not significantly different from the predicted repost count in the absence of notes (mean = 3146; $KS = 0.022$, $p = 0.310$; see Fig. S6.15). Note that the lack of statistical significance is likely partially due to the smaller sample size (only 3880 posts during this timeframe) and limited statistical power. However, the estimate aligns very closely with the 7.6% reduction reported in Chuai et al. (2025), which also did not reach statistical significance at common statistical significance thresholds.

Taken together, these findings demonstrate that both studies are consistent despite differences in methodology and timeframe: (i) in both timeframes, community notes had a substantial treatment effect at the individual post level, though the effect was smaller in the program's early stages. (ii) However, because time-to-display was significantly longer, the system-level impact was much smaller and not statistically significant in the

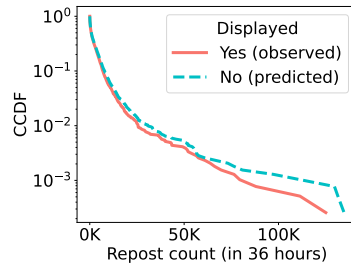


Figure S6.15: CCDFs showing the actually observed repost count for source posts with displayed community notes and the predicted repost count that the source posts would have received in the absence of community notes display. Only posts created from the roll-out of Community Notes program until April 2023 are included.

early stage of the program. Thus, rather than being contradictory, the findings from [74] and this study offer complementary perspectives on how the system-level effect of community notes evolved as the feature improved.

S6.9.5 Alternative Before-Display Periods

In the main analysis, we omitted a small subset of community notes (5.2%) that were displayed within the initial four hours after post creation and used a four-hour before-display period for DiD estimation. This approach allowed us to incorporate sufficient pre-trend periods while minimizing the exclusion of early-displayed notes. However, we also found that community notes were more effective when displayed earlier. Given this, we conducted additional analyses based on shorter before-display periods to assess the extent to which this exclusion affected our estimates.

Table S6.52 reports the two-period DiD estimation results using a one-hour before-display period (Column (1)) and a two-hour before-display period (Column (2)). The ATT estimates were -0.629 (99% CI: $[-0.636, -0.623]$; $p < 0.001$) based on the one-hour before-display period and -0.628 (99% CI: $[-0.633, -0.623]$; $p < 0.001$) based on the two-hour before-display period. Compared to our main ATT estimate based on the four-hour pre-display period (ATT of -0.612), these values differed by only 0.017 and 0.016, respectively. These minimal differences imply that the exclusion of early-displayed notes had no meaningful effect on our estimates.

S6.9.6 Analysis With HonestDiD

To further validate the robustness of our findings, we quantified the aggregated average treatment effect on the treated (ATT) after the display of Community Notes using the HonestDiD method. Specifically, we performed a multi-period ATT estimation, treating the entire post-display period as a single interval, while the four hourly periods preceding the display were used to assess pre-trends within the HonestDiD framework.

The full estimation results are reported in Table S6.53, where *After* indicates the whole after-display period from 1 to 12 hours, and the hourly after-display period *After : 12* is omitted due to collinearity. The aggregated ATT estimate was -0.600 (99% CI: $[-0.639, -0.557]$) when allowing after-display deviation to be twice as large as the maximum before-display deviation ($\bar{M} = 2$, as shown in Fig. S6.16), which is consistent with our main two-period ATT estimate. Furthermore, based on the robust HonestDiD estimate, the predicted reduction in cumulative

Table S6.51: Regression results for the treatment effect estimation with the restrictions on note scores and pre-display engagement (Column (1)–(2)) and the early stage of Community Notes program (Column (3)). Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

	Restriction on note scores and pre-display engagement (1) DiD	(2) RDD	Early stage of the program (3) RDD
Display	0.106 (0.102)		1.342*** (0.089)
Helpful		−0.006 (0.257)	
After	0.225*** (0.020)	0.267*** (0.023)	0.162*** (0.014)
Display×After	−0.905*** (0.021)		−0.675*** (0.015)
Helpful×After		−0.958*** (0.025)	
NoteScore		2.372 (4.081)	
PostAge	−0.793*** (0.012)	−0.833*** (0.014)	−0.811*** (0.008)
Intercept	2.213*** (0.073)	1.133 (1.510)	1.295*** (0.063)
Post-level RE	✓	✓	✓
#Observations	26,016	19,056	49,184
#Posts	1626	1191	3074

Table S6.52: Regression results for two-period ATT estimation with a before-display period of one hour (Column (1)) and a before-display period of two hours (Column (2)). Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

	(1) One-hour before-display period	(2) Two-hour before-display period
Display	1.258*** (0.023)	1.246*** (0.023)
After	0.066*** (0.005)	0.070*** (0.004)
Display×After	−0.992*** (0.007)	−0.988*** (0.005)
PostAge	−0.787*** (0.002)	−0.790*** (0.002)
Intercept	1.193*** (0.017)	1.255*** (0.016)
Post-level RE	✓	✓
#Observations	580,112	623,000
#Posts	44,624	44,500

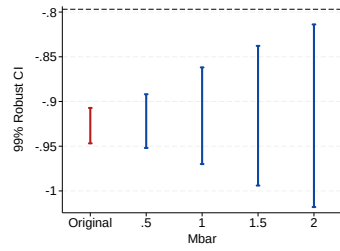


Figure S6.16: The original DiD coefficient estimate for the aggregated ATT after the display of community notes and its robust confidence intervals under different $\bar{M}$ from 0.5 to 2 with an increment of 0.5.

repost count is, on average, 409, indicating that Community Notes reduced the overall number of reposts on $\mathbb{X}$ by 14.4% (median of 9.7%). This is very close to the estimated effect on cumulative repost count in the main analysis (14.9%).

Table S6.53: Regression results for multi-period ATTs estimation, treating the entire after-display period as a single interval. Post-specific random effects are included. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

Multi-period			
Display	1.172 *** (0.024)		
Before:4	0.160 *** (0.010)		
Before:3	0.054 *** (0.009)		
Before:2	0.014 (0.008)		
After	−0.100 *** (0.025)		
After:1	0.367 *** (0.022)		
After:2	0.163 *** (0.021)		
After:3	0.075 *** (0.019)		
After:4	0.044* (0.017)		
After:5	0.016 (0.015)		
After:6	0.007 (0.013)		
After:7	−0.009 (0.011)		
After:8	−0.007 (0.010)		
After:9	−0.010 (0.008)		
After:10	−0.012 (0.007)		
After:11	−0.009 (0.006)		
After:12	Omitted		
		Continued	
		Display×Before:4	0.043 *** (0.010)
		Display×Before:3	0.044 *** (0.010)
		Display×Before:2	0.027* (0.010)
		Display×After	−0.927 *** (0.008)
		PostAge	−0.604 *** (0.015)
		Intercept	1.397 *** (0.019)
		Post-level RE	✓
		#Observations	654,400
		#Posts	40,900

S6.10 Analysis of Deleted and Suspended/Protected Posts

When using X API to retrieve posts fact-checked via community notes, we received return errors for 88,046 posts. Specifically, 64,792 posts were missing due to “not found error”, and 23,254 posts were missing due to “authorization error.” As detailed in the X API documentation (<https://developer.x.com/en/support/x-api/error-troubleshooting>), a “not found error” indicates that the corresponding post was deleted by the author of the post. An “authorization error” indicates that the corresponding post was protected by its poster (*i.e.*, private posts) or suspended by the X platform. Since the X API does not distinguish between these two scenarios, we conducted a manual review of 1,500 of these posts. Here, we found that almost all (94.4%) of these inaccessible posts resulted from account suspensions.

S6.10.1 Analysis of Deleted Posts

Regression Discontinuity Design (RDD) is a quasi-experimental research design that aims to estimate the effect of a treatment by exploiting a discontinuity in the relationship between a running variable and an outcome variable at a specific threshold or cutoff point [74, 184]. We used note helpfulness score (*NoteScore*) as a running variable to examine the changes in the probability of post deletion within a narrow window based on a RDD model. The “Community Notes” program adopted several models to determine the helpfulness of community notes, while the Core model was always authoritative [74]. To maintain consistency, we only considered note scores that were decided by the Core Model. The cut-off point of note helpfulness score was 0.40, and community notes that got score of 0.40 or above were displayed on the corresponding misleading posts. We considered the window of note helpfulness scores between 0.35 and 0.45. Additionally, due to the dynamic of note writing and rating, the recalculated note helpfulness scores may be fluctuate around the cut-off point and not exactly the same with that in the production. Given this, we excluded posts that have community notes with helpfulness score between 0.39 and 0.40 (*i.e.*, [0.39, 0.40)).

Subsequently, we specified our RDD model based on a logistic regression:

$$\text{logit}(\text{Deletion}_i) = \beta_0 + \beta_1 \text{Display}_i + \beta_2 \text{NoteScore}_i, \quad (\text{S6.3})$$

where *Display* indicate whether the note helpfulness score was ≥ 0.40 ($=1$) or not ($=0$), and *NoteScore* represented the running variable of the note helpfulness score. *NoteScore* was recentered around the cut-off point. The estimation results were reported in Column (1) of Table S6.54. The coefficient estimate for *Display* was 0.664 (99% CI: [0.477, 0.851]; $p < 0.001$). We further examined the treatment effect of community notes display on the post deletion based on the odds ratio estimated for *Display*: $e^\beta - 1$, where β denoted the coefficient estimate of *Display*. Specifically, the estimated treatment effect was 0.943 (99% CI: [0.611, 1.342]; $p < 0.001$). This indicates that posts with displayed community notes were 94.3% more likely to be deleted compared to posts without displayed community notes.

Additionally, we incorporated the interaction between *Display* and *NoteScore* in the RDD model to consider the potential different slope over note helpfulness scores after the cut-off point compared to that

Table S6.54: Estimation results for RDD logistic regression models predicting deletion of posts. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

	(1) Main	(2) Interaction
Display	0.664*** (0.073)	0.750*** (0.090)
NoteScore	4.681*** (1.143)	1.164 (2.424)
Display×NoteScore		4.496 (2.750)
Intercept	−2.268*** (0.045)	−2.379*** (0.082)
#Posts	36,135	36,135

Table S6.55: Estimation results for RDD logistic regression models predicting suspension/protection of posts. Reported are coefficient estimates with standard errors in parentheses. * $p < 0.01$, ** $p < 0.005$, *** $p < 0.001$.

	(1) Main	(2) Interaction
Display	0.072 (0.128)	0.238 (0.154)
NoteScore	5.646* (2.064)	−2.130 (3.976)
Display×NoteScore		10.792 (4.662)
Intercept	−3.279*** (0.077)	−3.524*** (0.135)
#Posts	36,135	36,135

before the cut-off point. The estimation results are reported in Column (2) of Table S6.54. The coefficient estimate of *NoteScore* was 1.164 (99% CI: [−5.080, 7.409]; $p = 0.631$) and not statistically significant. This indicates that the probability of post deletion remained relatively stable around the cut-off point in the absence of community notes. Moreover, the coefficient estimate for *Display* × *NoteScore* was 4.496 (99% CI: [−2.587, 11.579]; $p = 0.102$) and not statistically significant. This suggests that community notes had a sharp effect on post deletion once they were displayed on the misleading posts.

S6.10.2 Analysis of Protected/Suspended Posts

We applied the same RDD approach used in the analysis of deleted posts to examine the potential effect of community notes on post suspension/protection. The full estimation results are reported in Table S6.55. The coefficient estimates for *Display* were not statistically significant in both the Main model (*coef.* = 0.072; 99% CI: [−0.258, 0.401]; $p = 0.575$) and the Interaction model (*coef.* = 0.238; 99% CI: [−0.158, 0.634]; $p = 0.121$). This indicates that the display of community notes had no statistically significant treatment effect on post suspension/protection.

S6.11 Changes in X’s Algorithms and Policies on Misinformation

Our dataset covers all community fact-checked posts written in English between the roll-out of “Community Notes” to the general public on

October 6, 2022, and June 11, 2024 (*i.e.*, more than 20 months). During our observation period, X/Twitter has made several updates to its algorithms and its policies on misinformation. In the following, we provide an overview of the most relevant changes (Suppl. S6.11.1); and detail the steps we took to ensure that our empirical analysis accounts for these changes (Suppl. S6.11.2).

S6.11.1 Timeline of Changes

October 2022: Roll-out of the community notes program to the general public in the U.S. Users who are part of the program can add notes to tweets to add context and users can then vote if they determine the context to be helpful.

December 2022:

- ▶ Community notes become visible around the world. Prior to this global expansion, Community Notes were only visible to users in the U.S..
- ▶ Twitter disperses the Trust & Safety Council.
- ▶ Twitter updated its private information and media policy to prohibit sharing live location information.
- ▶ According to posts on Twitter from a former public policy employee, Twitter cut half of its public policy team.

April 2023:

- ▶ Twitter label tweets that get downranked for violating its hate speech policy.
- ▶ Twitter now shows labels on tweets with reduced visibility.
- ▶ Twitter updated its content moderation guidelines regarding hateful content, removing a policy that prohibited the targeted deadnaming or misgendering of transgender people.
- ▶ EU warns Twitter over disinformation

May 2023: Twitter launched Community Notes for images in posts.

August 2023: X changes the presentation format of Community Notes. The feature historically provided information to users about why they were seeing a fact check appended to a tweet. After three years of development, the company says this extra context will now be removed for those who are already experienced with Community Notes.

September 2023: Community Notes is now available for videos on X. X announced that notes by contributors attached to a video will show up in all posts with that video.

October 2023: A series of updates were made to Community Notes after the fact-checking system took multiple days to correct misinformation related to the Israel-Hamas war. The changes included:

- ▶ X's CEO Linda Yaccarino promised Community Notes would "appear more quickly on X"
- ▶ Note previews are now supported on both Android and the web, with support "coming soon" to iOS
- ▶ Improved media matching
- ▶ Notifications will send to users who Liked, Reposted or Replied to a post that later received a note

January 2024: X announces plans to hire 100 moderators in Austin.

June 2024: Adult content and violent content policies replace the former sensitive media and violent speech policies.

S6.11.2 Accounting for Algorithmic Effects in Our Study

We carefully accounted for potential algorithmic effects across all aspects of our study. This included (i) reviewing X's recommendation algorithm, (ii) employing robust model specifications, (iii) conducting comprehensive sensitivity analyses, and (iv) differentiating between user-initiated and platform-enforced actions. Together, these measures ensure that the observed changes in resharing behavior after a Community Note becomes visible are driven by shifts in user behavior rather than platform interventions.

- **Review of X's recommendation algorithm:** X's FAQs state that *"Notes that have been rated helpful by enough contributors from different points of view will appear directly on posts. Beyond that, notes do not affect display of posts or enforcement of X's rules"* (<https://communitynotes.x.com/guide/en/about/faq>). Furthermore, X's recommendation algorithm is available as open source (<https://github.com/twitter/the-algorithm>). Consistent with X's FAQs, the algorithm does not impose any penalties (e.g., visibility reduction) or trigger enforcement of X's rules (e.g., account suspensions) on posts flagged with community notes. Consequently, any observed changes in resharing behavior after a community note becomes visible must be attributed to shifts in user behavior, rather than algorithmic intervention.
- **Model specification:** Our model explicitly accounts for algorithmic changes by allowing the treatment effect to vary over time. To achieve this, all our models incorporate a comprehensive set of control variables as well as post-specific random effects (see Section 6.4). Additionally, our propensity score matching method ensures that we compare only posts with and without displayed notes that were published within a similar timeframe (i.e., during the same month).
- **Analysis of treatment effect over time:** As part of our sensitivity analyses, we specifically examined how the efficacy of Community Notes has evolved since the feature's rollout in October 2022 (see Section 6.2). Our analysis revealed a trend of increasing efficacy over time, which may be partly attributed to improvements within the Community Notes program, such as faster note generation.
- **Analysis of suspensions:** We investigated the potential impact of Community Notes on account suspensions (see Section 6.2). Unlike post deletions (i.e., user-initiated actions), we observed no significant change in the ratio of suspended posts (i.e., platform-enforced actions).

S7

Supplementary Materials for Chapter 7

S7.1 Data Overview

S7.1.1 Correlations Between Sentiments/Emotions in Source Posts and Sentiments/Emotions in Replies

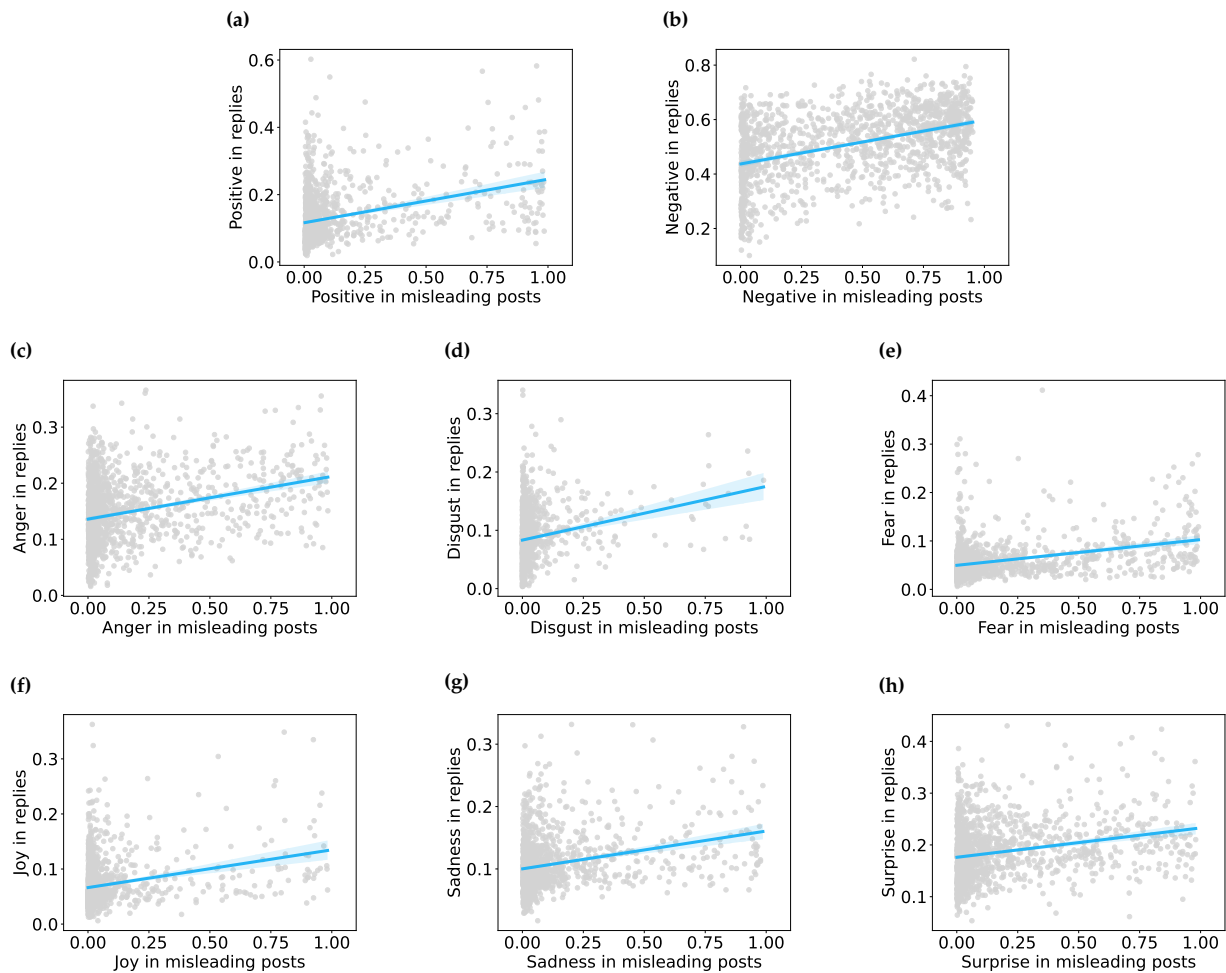


Figure S7.1: The Pearson correlations between sentiments (emotions) in replies and source sentiments (emotions) in misleading posts. (a) Positive sentiment ($r = 0.365$, $p < 0.001$). (b) Negative sentiment ($r = 0.426$, $p < 0.001$). (c) Anger ($r = 0.316$, $p < 0.001$). (d) Disgust ($r = 0.235$, $p < 0.001$). (e) Fear ($r = 0.354$, $p < 0.001$). (f) Joy ($r = 0.287$, $p < 0.001$). (g) Sadness ($r = 0.302$, $p < 0.001$). (h) Surprise ($r = 0.254$, $p < 0.001$).

Given the potential emotion transfers from source posts to their replies, we analyze the correlations between source sentiments and emotions in the misleading posts and the subsequent sentiments and emotions in the replies. Of note, the replies after the display of community notes

are excluded in the analysis of emotion transfers to prevent the potential contamination of note display. We find that all the sentiments and emotions in replies are positively correlated to the corresponding source sentiments and emotions in the misleading posts (see Fig. S7.1).

S7.1.2 Sentiments and Emotions in Replies Over Time

Pre-trends before note display. We analyze whether the sentiments and emotions in replies change following the spread of source posts (*i.e.*, post age) before the display of community notes. Specifically, we take sentiments and emotions in replies before note display as dependent variables and examine the effect of the independent variable of *PostAge*. The estimation results are reported in Table S7.1 and Table S7.2. All the coefficient estimates of *PostAge* are consistently not statistically significant across sentiments and emotions in replies before note display (each $p > 0.05$). This suggests that the sentiments and emotions in replies remain stable over time following the spread of posts.

Table S7.1: Estimation results for positive [Column (1)] and negative [Column (2)] sentiments in replies before note display. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) Positive	(2) Negative
SourcePositive	0.039* (0.017)	0.015 (0.020)
SourceNegative	−0.054*** (0.016)	0.122*** (0.022)
PostAge	−0.001 (0.006)	−0.004 (0.007)
Intercept	0.000 (0.013)	0.000 (0.018)
R^2	0.007	0.013
#Replies (N)	1,688,102	1,688,102
#Source posts	1840	1840

Discontinuity around the display of community notes. We use an interrupted time series design to detect discontinuity around the cut-off point indicated by the display of community notes. The estimation results for sentiments and emotions in replies within the entire lifespan are reported in Table S7.3 and Table S7.4. The key variables are *Displayed* and its interaction with *PostAge*. We find that the coefficient estimates of *Displayed* are significantly significant for positive, negative, anger, and surprise, indicating significant changes after the display of community notes. Additionally, the coefficient estimates of *Displayed* $\times$ *PostAge* are not statistically significant across sentiments and emotions in replies (each $p > 0.05$). This suggests that the sentiments and emotions in replies remain stable over time after the changes at the cut-off point. Moreover, the coefficient estimates of *PostAge* are consistent with that in the analysis of pre-trends for sentiments and emotions in replies (each $p > 0.05$). Taken together, these findings suggest the existence of discontinuities for certain sentiments and emotions at the cut-off point.

To better clarify the stable trends in sentiments and emotions over time and discontinuities at the cut-off point, we illustrate the hourly averages of sentiments and emotions in replies within the 16-hour bandwidth (see Fig. S7.2).

Table S7.2: Estimation results for anger [Column (1)], disgust [Column (2)], fear [Column (3)], joy [Column (4)], sadness [Column (5)], and surprise [Column (6)] in replies before note display. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) Anger	(2) Disgust	(3) Fear	(4) Joy	(5) Sadness	(6) Surprise
SourceAnger	0.083*** (0.014)	0.010 (0.011)	0.017 (0.009)	-0.005 (0.009)	0.019 (0.015)	-0.039*** (0.011)
SourceDisgust	0.000 (0.008)	0.073*** (0.015)	-0.004 (0.004)	-0.009 (0.005)	-0.010 (0.006)	-0.004 (0.007)
SourceFear	0.013 (0.015)	0.011 (0.012)	0.067*** (0.011)	-0.023** (0.009)	0.003 (0.010)	0.002 (0.014)
SourceJoy	0.006 (0.014)	0.000 (0.009)	0.005 (0.006)	0.038** (0.014)	0.024** (0.008)	0.003 (0.011)
SourceSadness	0.009 (0.010)	0.001 (0.010)	0.005 (0.005)	0.010 (0.009)	0.058*** (0.009)	-0.012 (0.009)
SourceSurprise	-0.054*** (0.011)	-0.005 (0.013)	-0.001 (0.007)	0.000 (0.010)	0.012 (0.013)	0.065*** (0.013)
PostAge	-0.004 (0.004)	0.011 (0.009)	0.004 (0.004)	0.001 (0.005)	0.000 (0.003)	0.006 (0.004)
Intercept	0.000 (0.012)	0.000 (0.010)	0.000 (0.008)	0.000 (0.009)	0.000 (0.012)	0.000 (0.010)
R^2	0.013	0.006	0.005	0.003	0.003	0.008
#Replies (N)	1,688,102	1,688,102	1,688,102	1,688,102	1,688,102	1,688,102
#Source posts	1840	1840	1840	1840	1840	1840

	(1) Positive	(2) Negative
Displayed	-0.057** (0.018)	0.117 *** (0.024)
PostAge	-0.004 (0.013)	-0.009 (0.015)
Displayed × PostAge	0.008 (0.013)	0.004 (0.014)
SourcePositive	0.039* (0.018)	0.015 (0.020)
SourceNegative	-0.047** (0.016)	0.113 *** (0.023)
Intercept	0.013 (0.013)	-0.029 (0.018)
R^2	0.006	0.014
#Replies (N)	2,225,260	2,225,260
#Source posts	1841	1841

Table S7.3: Estimation results for positive [Column (1)] and negative [Column (2)] sentiments in replies. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table S7.4: Estimation results for anger [Column (1)], disgust [Column (2)], fear [Column (3)], joy [Column (4)], sadness [Column (5)], and surprise [Column (6)] in replies. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) Anger	(2) Disgust	(3) Fear	(4) Joy	(5) Sadness	(6) Surprise
Displayed	0.080*** (0.012)	0.021 (0.011)	−0.008 (0.009)	−0.017 (0.011)	−0.011 (0.011)	−0.049*** (0.010)
PostAge	−0.009 (0.009)	0.025 (0.019)	0.009 (0.010)	0.001 (0.010)	−0.001 (0.008)	0.014 (0.010)
Displayed × PostAge	0.006 (0.008)	−0.028 (0.019)	−0.006 (0.009)	0.004 (0.010)	0.005 (0.008)	−0.008 (0.010)
SourceAnger	0.078*** (0.012)	0.008 (0.011)	0.016* (0.008)	−0.006 (0.008)	0.020 (0.014)	−0.037*** (0.010)
SourceDisgust	0.001 (0.008)	0.069*** (0.015)	−0.004 (0.004)	−0.011* (0.005)	−0.012* (0.006)	−0.004 (0.007)
SourceFear	0.013 (0.014)	0.008 (0.011)	0.066*** (0.010)	−0.021** (0.007)	0.006 (0.009)	0.004 (0.012)
SourceJoy	0.009 (0.016)	−0.001 (0.009)	0.005 (0.006)	0.032 (0.018)	0.026*** (0.007)	0.000 (0.012)
SourceSadness	0.008 (0.010)	−0.002 (0.009)	0.006 (0.005)	0.009 (0.009)	0.056*** (0.008)	−0.008 (0.008)
SourceSurprise	−0.055*** (0.011)	−0.012 (0.012)	−0.002 (0.007)	0.000 (0.009)	0.016 (0.012)	0.068*** (0.013)
Intercept	−0.020 (0.011)	−0.002 (0.010)	0.002 (0.008)	0.004 (0.009)	0.002 (0.012)	0.012 (0.010)
R^2	0.014	0.006	0.005	0.002	0.003	0.009
#Replies (N)	2,225,260	2,225,260	2,225,260	2,225,260	2,225,260	2,225,260
#Source posts	1841	1841	1841	1841	1841	1841

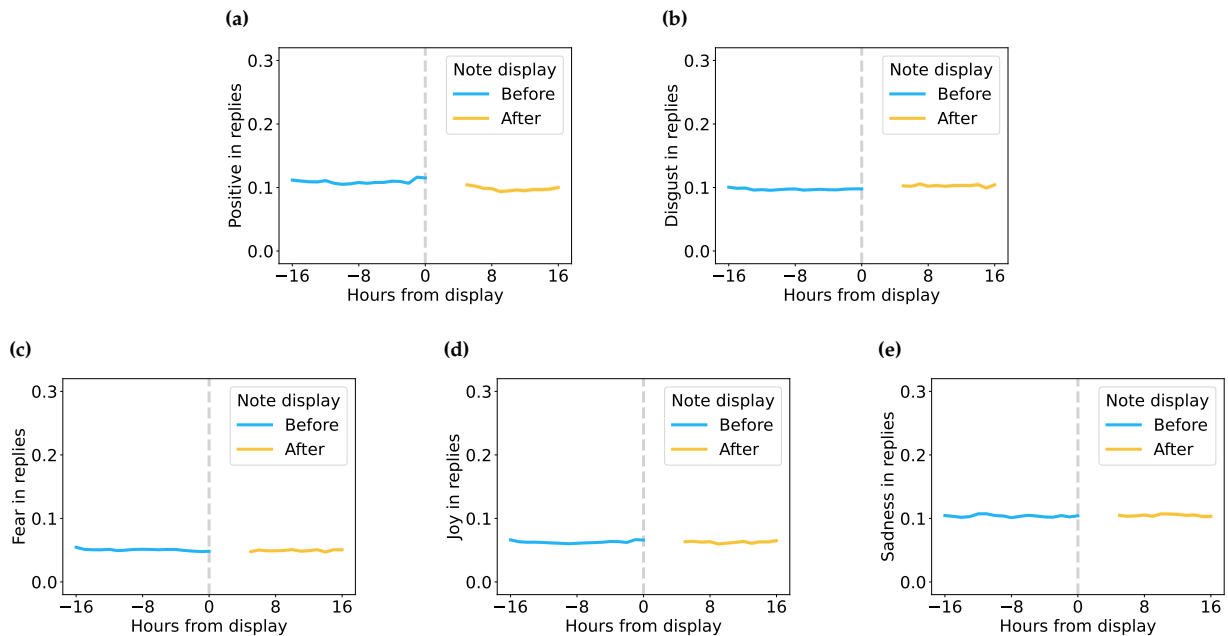


Figure S7.2: The hourly averages in sentiments and emotions across hours from note display. (a) The hourly averages of positive sentiment in replies across hours from note display. (b) The hourly averages of disgust in replies across hours from note display. (c) The hourly averages of fear in replies across hours from note display. (d) The hourly averages of joy in replies across hours from note display. (e) The hourly averages of sadness in replies across hours from note display. The error bands represent 95% CIs.

S7.1.3 Correlations Between Sentiments and Emotions

The Pearson correlations between sentiments and emotions are shown in Table S7.5. Anger, disgust, fear, and sadness are positively associated with negative sentiment, while joy and surprise are positively associated with positive sentiment.

Table S7.5: The Pearson correlations between sentiments and emotions with two-sided p-values. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	Anger	Disgust	Fear	Joy	Sadness	Surprise
Positive	−0.204***	−0.175***	−0.074***	0.572***	−0.119***	0.044***
Negative	0.406***	0.323***	0.089***	−0.324***	0.192***	−0.141***

S7.1.4 Correlations Among Emotions

The Pearson correlations among the six basic emotions in replies are reported in Table S7.6. Only anger and disgust are positively correlated, and all the other pairs of emotions are negatively correlated.

Table S7.6: The Pearson correlations among emotions with two-sided p-values. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	Anger	Disgust	Fear	Joy	Sadness	Surprise
Anger						
Disgust	0.111***					
Fear	−0.040***	−0.076***				
Joy	−0.195***	−0.168***	−0.085***			
Sadness	−0.147***	−0.129***	−0.062***	−0.095***		
Surprise	−0.313***	−0.241***	−0.136***	−0.066***	−0.131***	

S7.2 Descriptive Analysis

Fig. S7.3 shows the comparisons between means before note display and means after note display across sentiments and emotions. For sentiments, positive sentiment in replies after the display of community notes (mean of 0.115) is significantly lower than before the note display (mean of 0.127; $t = 10.249, p < 0.001$; Fig. S7.3a). In contrast, negative sentiment in replies after the display of community notes (mean of 0.566) is significantly higher than before the note display (mean of 0.524; $t = -23.748, p < 0.001$; Fig. S7.3b). In terms of basic emotions, anger (mean of 0.179) and disgust (mean of 0.093) in replies after the display of community notes are significantly higher than anger (mean of 0.155; $t = -18.531, p < 0.001$; Fig. S7.3c) and disgust (mean of 0.090; $t = -2.717, p < 0.01$; Fig. S7.3d) in replies before the note display, respectively. Fear (before: 0.057, after: 0.057; $t = 0.025, p = 0.980$; Fig. S7.3e), joy (before: 0.071, after: 0.070; $t = 1.781, p = 0.075$; Fig. S7.3f), and sadness (before: 0.110, after: 0.110; $t = -0.745, p = 0.456$; Fig. S7.3g) have no statistically significant changes after the display of community notes. Surprise in replies after the display of community notes (mean of 0.179) is significantly lower than that in replies before the note display (mean of 0.185; $t = 5.325, p < 0.001$; Fig. S7.3h).

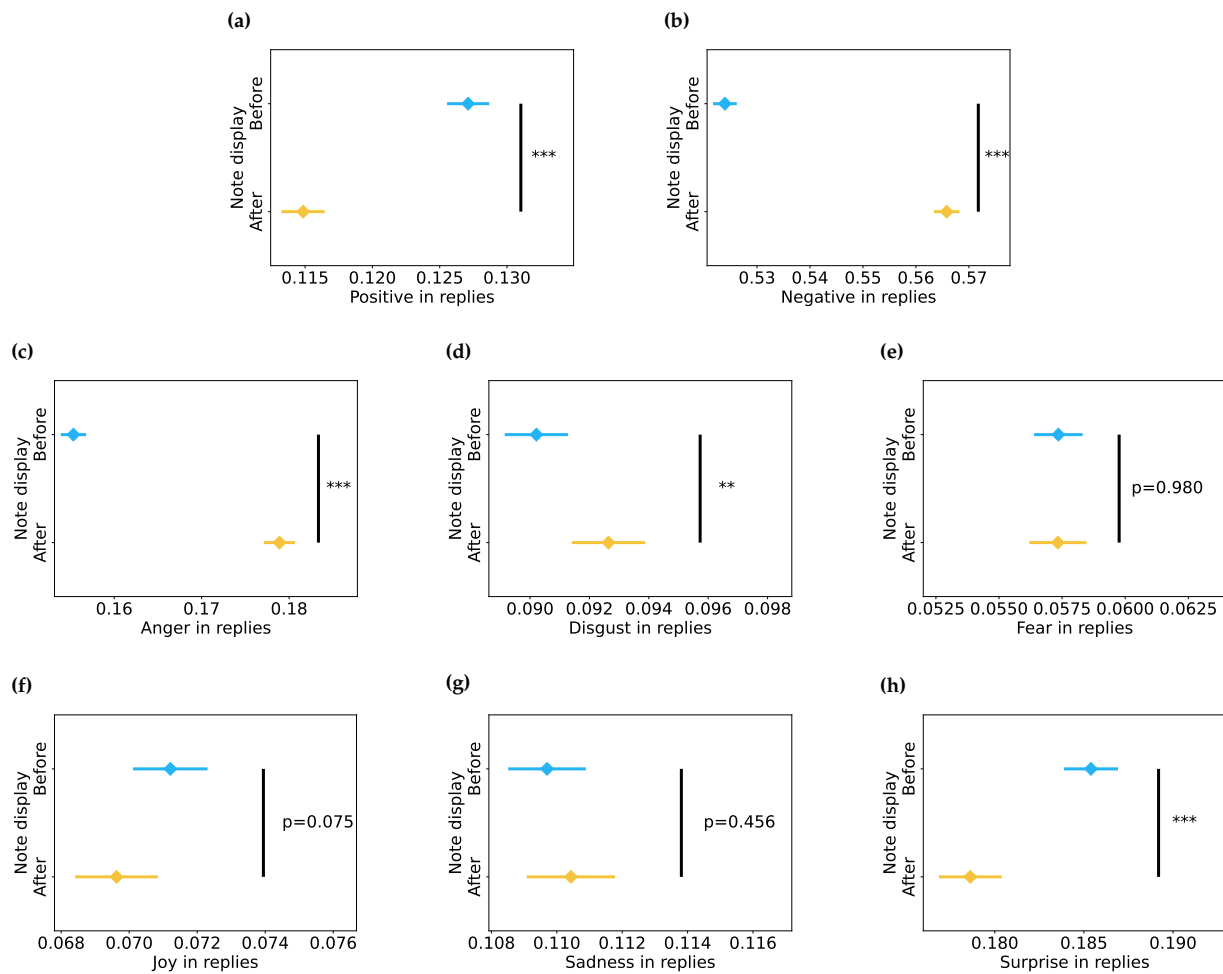


Figure S7.3: The sentiments and emotions in replies before and after the display of community notes. Shown are mean values with error bars representing 95% CIs. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. (a) Positive sentiment in replies. (b) Negative sentiment in replies. (c) Anger in replies. (d) Disgust in replies. (e) Fear in replies. (f) Joy in replies. (g) Sadness in replies. (h) Surprise in replies.

S7.3 Estimation Results for Main Analysis

The estimation results for the main analysis on the changes in specific sentiments and emotions in replies after the display of community notes are reported in Table S7.7, Table S7.8, and Table S7.9. Here, we use all direct replies within the entire lifespan for coefficient estimation. Subsequently, we estimate the effect sizes of community notes on sentiments and emotions within a 16-hour bandwidth (see Fig. 7.4). The equivalence of RDD models across short and long bandwidths is validated in Suppl. S7.4.

	(1) Positive	(2) Negative
Displayed	−0.055** (0.017)	0.114 *** (0.022)
HoursFromDisplay	−0.003 (0.014)	−0.001 (0.021)
SourcePositive	0.039* (0.019)	0.015 (0.021)
SourceNegative	−0.048** (0.017)	0.112 *** (0.023)
PostAge	0.005 (0.009)	−0.005 (0.013)
Intercept	0.014 (0.014)	−0.028 (0.018)
R^2	0.007	0.013
#Replies (N)	2,155,873	2,155,873
#Source posts	1339	1339

Table S7.7: Estimation results for positive sentiment [Column (1)] and negative sentiment [Column (2)] in replies within reply lifespan. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table S7.8: Estimation results for anger [Column (1)], disgust [Column (2)], and fear [Column (3)] in replies within reply lifespan. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) Anger	(2) Disgust	(3) Fear
Displayed	0.078*** (0.010)	0.023* (0.010)	−0.007 (0.008)
HoursFromDisplay	−0.004 (0.018)	−0.006 (0.010)	−0.001 (0.005)
SourceAnger	0.078*** (0.012)	0.007 (0.011)	0.017* (0.008)
SourceDisgust	0.001 (0.008)	0.070*** (0.015)	−0.004 (0.004)
SourceFear	0.013 (0.015)	0.008 (0.011)	0.066*** (0.010)
SourceJoy	0.010 (0.016)	0.000 (0.009)	0.004 (0.006)
SourceSadness	0.010 (0.010)	−0.005 (0.009)	0.005 (0.005)
SourceSurprise	−0.056*** (0.011)	−0.013 (0.012)	−0.001 (0.007)
PostAge	−0.002 (0.010)	0.004 (0.008)	0.004 (0.004)
Intercept	−0.019 (0.012)	−0.006 (0.010)	0.002 (0.008)
R^2	0.014	0.006	0.005
#Replies (N)	2,155,873	2,155,873	2,155,873
#Source posts	1339	1339	1339

Table S7.9: Estimation results for joy [Column (1)], sadness [Column (2)], and surprise [Column (3)] in replies within reply lifespan. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) Joy	(2) Sadness	(3) Surprise
Displayed	−0.015 (0.010)	−0.008 (0.012)	−0.043*** (0.009)
HoursFromDisplay	−0.004 (0.010)	−0.004 (0.007)	−0.005 (0.010)
SourceAnger	−0.006 (0.008)	0.021 (0.014)	−0.036*** (0.010)
SourceDisgust	−0.011* (0.005)	−0.012 (0.006)	−0.003 (0.007)
SourceFear	−0.021** (0.008)	0.006 (0.009)	0.004 (0.012)
SourceJoy	0.033 (0.019)	0.025*** (0.007)	−0.001 (0.012)
SourceSadness	0.009 (0.009)	0.052*** (0.008)	−0.007 (0.009)
SourceSurprise	0.000 (0.010)	0.016 (0.013)	0.069*** (0.013)
PostAge	0.007 (0.006)	0.006 (0.004)	0.010 (0.006)
Intercept	0.004 (0.009)	0.002 (0.012)	0.011 (0.010)
R^2	0.002	0.003	0.008
#Replies (N)	2,155,873	2,155,873	2,155,873
#Source posts	1339	1339	1339

S7.4 RDD Estimates Across Different Bandwidths

We examine the robustness of RDD estimates (*i.e.*, the coefficient estimates of *Displayed*) across three progressively extended bandwidths: 16 hours, one week, and the entire reply lifespan. Tables S7.10–S7.11 report the estimation results within 16-hour bandwidth, Tables S7.12–S7.13 report the estimation results within one-week bandwidth, and Tables S7.14–S7.15 report the estimation results within the entire reply lifespan. As shown in Fig. 7.4, the initial four hours after the display of community notes are omitted in the three cases to avoid potential cold-start contamination in the smaller bandwidth, such as the 16-hour window. We find that the RDD estimates across the three bandwidths are consistent, with no statistically significant differences observed.

	(1) Positive	(2) Negative
Displayed	−0.049* (0.024)	0.096*** (0.028)
HoursFromDisplay	−0.037 (0.127)	0.145 (0.163)
SourcePositive	0.053* (0.025)	0.000 (0.023)
SourceNegative	−0.030 (0.018)	0.093*** (0.023)
PostAge	0.008 (0.013)	−0.007 (0.018)
Intercept	−0.002 (0.016)	−0.031 (0.022)
R^2	0.006	0.012
#Replies (N)	1,166,861	1,166,861
#Source posts	1335	1335

Table S7.10: Estimation results for positive sentiment [Column (1)] and negative sentiment [Column (2)] in replies within 16-hour bandwidth. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table S7.11: Estimation results for anger [Column (1)], disgust [Column (2)], fear [Column (3)], joy [Column (4)], sadness [Column (5)], and surprise [Column (6)] in replies within 16-hour bandwidth. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) Anger	(2) Disgust	(3) Fear	(4) Joy	(5) Sadness	(6) Surprise
Displayed	0.059** (0.018)	0.039* (0.017)	0.014 (0.012)	−0.011 (0.020)	0.016 (0.014)	−0.059*** (0.014)
HoursFromDisplay	0.087 (0.092)	−0.084 (0.078)	−0.080 (0.059)	0.001 (0.078)	−0.056 (0.076)	0.025 (0.066)
SourceAnger	0.068*** (0.013)	0.008 (0.011)	0.012 (0.007)	−0.007 (0.007)	0.018 (0.012)	−0.029** (0.009)
SourceDisgust	−0.004 (0.009)	0.074*** (0.017)	−0.007 (0.005)	−0.008 (0.006)	−0.009 (0.005)	0.000 (0.007)
SourceFear	0.017 (0.016)	0.010 (0.011)	0.064*** (0.011)	−0.023** (0.007)	0.008 (0.009)	0.000 (0.009)
SourceJoy	0.007 (0.020)	−0.001 (0.008)	0.004 (0.007)	0.028 (0.023)	0.023*** (0.007)	−0.002 (0.012)
SourceSadness	0.002 (0.011)	−0.004 (0.009)	0.006 (0.007)	0.007 (0.010)	0.046*** (0.008)	−0.001 (0.009)
SourceSurprise	−0.062*** (0.014)	−0.021 (0.012)	−0.003 (0.009)	−0.003 (0.010)	0.028 (0.016)	0.074*** (0.016)
PostAge	−0.011 (0.011)	0.020 (0.015)	0.008 (0.009)	0.009 (0.011)	0.007 (0.009)	0.010 (0.009)
Intercept	−0.014 (0.015)	0.008 (0.013)	0.008 (0.009)	−0.012 (0.011)	−0.010 (0.013)	0.000 (0.011)
R^2	0.014	0.008	0.004	0.002	0.002	0.009
#Replies (N)	1,166,861	1,166,861	1,166,861	1,166,861	1,166,861	1,166,861
#Source posts	1335	1335	1335	1335	1335	1335

Table S7.12: Estimation results for positive sentiment [Column (1)] and negative sentiment [Column (2)] in replies within one-week bandwidth. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) Positive	(2) Negative
Displayed	−0.040 (0.023)	0.113 *** (0.028)
HoursFromDisplay	−0.060 (0.035)	0.065 (0.043)
SourcePositive	0.036* (0.018)	0.016 (0.021)
SourceNegative	−0.047** (0.017)	0.110 *** (0.024)
PostAge	0.005 (0.011)	−0.016 (0.018)
Intercept	0.011 (0.013)	−0.027 (0.018)
R^2	0.007	0.014
#Replies (N)	1,952,474	1,952,474
#Source posts	1339	1339

Table S7.13: Estimation results for anger [Column (1)], disgust [Column (2)], fear [Column (3)], joy [Column (4)], sadness [Column (5)], and surprise [Column (6)] in replies within one-week bandwidth. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) Anger	(2) Disgust	(3) Fear	(4) Joy	(5) Sadness	(6) Surprise
Displayed	0.067*** (0.012)	0.035* (0.017)	−0.008 (0.008)	−0.001 (0.015)	0.010 (0.013)	−0.047*** (0.014)
HoursFromDisplay	0.066*** (0.020)	−0.033 (0.036)	0.000 (0.017)	−0.044 (0.024)	−0.033 (0.018)	−0.020 (0.028)
SourceAnger	0.069*** (0.011)	0.006 (0.011)	0.016 (0.009)	0.000 (0.008)	0.025 (0.016)	−0.033** (0.010)
SourceDisgust	0.002 (0.008)	0.072*** (0.015)	−0.004 (0.004)	−0.011* (0.005)	−0.011 (0.006)	−0.004 (0.007)
SourceFear	0.018 (0.015)	0.012 (0.011)	0.069*** (0.011)	−0.024** (0.007)	0.003 (0.009)	−0.002 (0.011)
SourceJoy	0.008 (0.016)	0.001 (0.008)	0.005 (0.007)	0.033 (0.018)	0.025*** (0.007)	0.000 (0.012)
SourceSadness	0.008 (0.010)	−0.004 (0.009)	0.005 (0.005)	0.010 (0.009)	0.053*** (0.009)	−0.007 (0.009)
SourceSurprise	−0.055*** (0.011)	−0.010 (0.012)	0.000 (0.007)	0.001 (0.010)	0.017 (0.013)	0.068*** (0.013)
PostAge	−0.018* (0.009)	0.019 (0.013)	0.009 (0.009)	0.007 (0.008)	0.009 (0.006)	0.016 (0.009)
Intercept	−0.019 (0.011)	−0.005 (0.010)	0.003 (0.008)	0.001 (0.009)	0.000 (0.013)	0.010 (0.010)
R^2	0.014	0.006	0.005	0.002	0.003	0.008
#Replies (N)	1,952,474	1,952,474	1,952,474	1,952,474	1,952,474	1,952,474
#Source posts	1339	1339	1339	1339	1339	1339

	(1) Positive	(2) Negative
Displayed	−0.071*** (0.021)	0.144*** (0.026)
HoursFromDisplay	−0.002 (0.014)	−0.002 (0.021)
SourcePositive	0.038* (0.018)	0.016 (0.021)
SourceNegative	−0.049** (0.017)	0.113*** (0.024)
PostAge	0.007 (0.009)	−0.008 (0.013)
Intercept	0.014 (0.014)	−0.029 (0.018)
R^2	0.007	0.014
#Replies (N)	2,000,658	2,000,658
#Source posts	1339	1339

Table S7.14: Estimation results for positive sentiment [Column (1)] and negative sentiment [Column (2)] in replies within reply lifespan. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table S7.15: Estimation results for anger [Column (1)], disgust [Column (2)], fear [Column (3)], joy [Column (4)], sadness [Column (5)], and surprise [Column (6)] in replies within reply lifespan. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) Anger	(2) Disgust	(3) Fear	(4) Joy	(5) Sadness	(6) Surprise
Displayed	0.096*** (0.011)	0.027* (0.011)	−0.004 (0.008)	−0.022 (0.012)	−0.004 (0.013)	−0.052*** (0.010)
HoursFromDisplay	−0.005 (0.018)	−0.006 (0.010)	−0.001 (0.005)	−0.004 (0.010)	−0.004 (0.007)	−0.005 (0.010)
SourceAnger	0.079*** (0.012)	0.006 (0.011)	0.017* (0.008)	−0.006 (0.008)	0.020 (0.015)	−0.037*** (0.010)
SourceDisgust	0.001 (0.008)	0.071*** (0.015)	−0.005 (0.004)	−0.010* (0.005)	−0.012 (0.006)	−0.003 (0.007)
SourceFear	0.013 (0.015)	0.009 (0.011)	0.066*** (0.011)	−0.021** (0.008)	0.005 (0.009)	0.004 (0.013)
SourceJoy	0.010 (0.016)	0.000 (0.009)	0.004 (0.006)	0.032 (0.018)	0.024** (0.007)	0.000 (0.012)
SourceSadness	0.010 (0.010)	−0.005 (0.009)	0.005 (0.005)	0.009 (0.009)	0.052*** (0.008)	−0.007 (0.009)
SourceSurprise	−0.056*** (0.011)	−0.011 (0.012)	−0.001 (0.007)	0.001 (0.010)	0.016 (0.013)	0.068*** (0.013)
PostAge	−0.004 (0.010)	0.004 (0.008)	0.004 (0.004)	0.008 (0.006)	0.006 (0.005)	0.011 (0.006)
Intercept	−0.020 (0.011)	−0.006 (0.010)	0.002 (0.008)	0.004 (0.009)	0.002 (0.012)	0.011 (0.010)
R^2	0.015	0.006	0.005	0.002	0.003	0.008
#Replies (N)	2,000,658	2,000,658	2,000,658	2,000,658	2,000,658	2,000,658
#Source posts	1339	1339	1339	1339	1339	1339

S7.5 Sensitivity Analysis

The estimation results for the sensitivity analysis across political and non-political misleading posts are reported in Table S7.16, Table S7.17, and Table S7.18.

Table S7.16: Estimation results for positive and negative sentiments in replies to political and non-political misleading source posts. Column (1) reports the estimation result for positive sentiment in replies to political misleading source posts. Column (2) reports the estimation result for positive sentiment in replies to non-political misleading source posts. Column (3) reports the estimation result for negative sentiment in replies to political misleading source posts. Column (4) reports the estimation result for negative sentiment in replies to non-political misleading source posts. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) Positive: politics	(2) Positive: non-politics	(3) Negative: politics	(4) Negative: non-politics
Displayed	−0.068** (0.024)	−0.020 (0.016)	0.116 *** (0.025)	0.075*** (0.017)
HoursFromDisplay	0.002 (0.020)	−0.006 (0.010)	−0.010 (0.027)	0.003 (0.013)
SourcePositive	−0.015 (0.016)	0.099*** (0.019)	0.020 (0.022)	−0.010 (0.024)
SourceNegative	−0.069** (0.023)	−0.036 (0.019)	0.090*** (0.027)	0.137*** (0.028)
PostAge	0.003 (0.012)	0.000 (0.008)	−0.001 (0.015)	0.005 (0.011)
Intercept	−0.044* (0.019)	0.070*** (0.016)	0.102*** (0.022)	−0.146*** (0.021)
R^2	0.005	0.013	0.010	0.021
#Replies (N)	1,037,340	1,118,533	1,037,340	1,118,533
#Source posts	379	960	379	960

Table S7.17: Estimation results for anger, disgust, and fear in replies to political and non-political misleading source posts. Column (1) reports the estimation result for anger in replies to political misleading source posts. Column (2) reports the estimation result for anger in replies to non-political misleading source posts. Column (3) reports the estimation result for disgust in replies to political misleading source posts. Column (4) reports the estimation result for disgust in replies to non-political misleading source posts. Column (5) reports the estimation result for fear in replies to political misleading source posts. Column (6) reports the estimation result for fear in replies to non-political misleading source posts. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	Anger		Disgust		Fear	
	(1) Politics	(2) Non-politics	(3) Politics	(4) Non-politics	(5) Politics	(6) Non-politics
Displayed	0.063*** (0.014)	0.070*** (0.015)	0.027** (0.010)	0.012 (0.016)	0.005 (0.009)	−0.015 (0.011)
HoursFromDisplay	−0.012 (0.029)	0.000 (0.005)	0.000 (0.011)	−0.011 (0.018)	0.001 (0.006)	−0.001 (0.005)
SourceAnger	0.075*** (0.015)	0.062*** (0.016)	0.018 (0.012)	−0.006 (0.017)	0.018* (0.008)	0.018 (0.015)
SourceDisgust	0.011 (0.012)	0.009 (0.009)	0.085*** (0.015)	0.064** (0.021)	−0.006 (0.009)	−0.005 (0.005)
SourceFear	0.019 (0.026)	0.013 (0.014)	0.007 (0.013)	0.010 (0.016)	0.042*** (0.011)	0.084*** (0.014)
SourceJoy	0.021 (0.015)	−0.014 (0.012)	0.016* (0.007)	−0.017 (0.012)	0.007 (0.008)	0.003 (0.009)
SourceSadness	0.022 (0.016)	0.003 (0.012)	0.017 (0.012)	−0.019 (0.012)	0.008 (0.010)	0.005 (0.006)
SourceSurprise	−0.051** (0.017)	−0.057*** (0.011)	0.002 (0.014)	−0.023 (0.018)	−0.016 (0.009)	0.008 (0.009)
PostAge	0.003 (0.017)	0.001 (0.004)	−0.004 (0.006)	0.010 (0.014)	−0.001 (0.004)	0.005 (0.005)
Intercept	0.065*** (0.017)	−0.096*** (0.011)	0.005 (0.012)	−0.017 (0.014)	−0.028** (0.009)	0.025 (0.013)
R^2	0.012	0.013	0.005	0.008	0.003	0.007
#Replies (N)	1,037,340	1,118,533	1,037,340	1,118,533	1,037,340	1,118,533
#Source posts	379	960	379	960	379	960

Table S7.18: Estimation results for joy, sadness, and surprise in replies to political and non-political misleading source posts. Column (1) reports the estimation result for joy in replies to political misleading source posts. Column (2) reports the estimation result for joy in replies to non-political misleading source posts. Column (3) reports the estimation result for sadness in replies to political misleading source posts. Column (4) reports the estimation result for sadness in replies to non-political misleading source posts. Column (5) reports the estimation result for surprise in replies to political misleading source posts. Column (6) reports the estimation result for surprise in replies to non-political misleading source posts. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	Joy		Sadness		Surprise	
	(1) Politics	(2) Non-politics	(3) Politics	(4) Non-politics	(5) Politics	(6) Non-politics
Displayed	−0.019 (0.014)	0.003 (0.011)	−0.014 (0.021)	−0.007 (0.011)	−0.025* (0.012)	−0.048*** (0.013)
HoursFromDisplay	−0.004 (0.017)	−0.005 (0.007)	−0.006 (0.016)	−0.002 (0.004)	−0.016 (0.020)	0.004 (0.007)
SourceAnger	−0.016 (0.011)	0.011 (0.011)	0.021 (0.021)	0.019 (0.012)	−0.050*** (0.013)	−0.017 (0.015)
SourceDisgust	−0.018* (0.008)	−0.011 (0.007)	0.003 (0.011)	−0.018** (0.007)	−0.014 (0.010)	−0.002 (0.008)
SourceFear	−0.017 (0.012)	−0.026** (0.009)	0.008 (0.014)	0.004 (0.011)	−0.002 (0.022)	0.007 (0.012)
SourceJoy	0.008 (0.016)	0.062*** (0.016)	0.033* (0.014)	0.017* (0.007)	−0.020 (0.012)	0.021 (0.013)
SourceSadness	−0.004 (0.013)	0.016 (0.011)	0.064*** (0.014)	0.044*** (0.011)	−0.023 (0.014)	0.002 (0.011)
SourceSurprise	−0.005 (0.014)	0.004 (0.012)	0.020 (0.013)	0.013 (0.020)	0.055*** (0.015)	0.078*** (0.018)
PostAge	0.007 (0.010)	0.005 (0.005)	0.007 (0.009)	0.006 (0.003)	0.017 (0.012)	0.003 (0.005)
Intercept	−0.020 (0.013)	0.029* (0.012)	0.013 (0.019)	−0.006 (0.014)	−0.022 (0.014)	0.042** (0.013)
R^2	0.001	0.005	0.003	0.003	0.009	0.008
#Replies (N)	1,037,340	1,118,533	1,037,340	1,118,533	1,037,340	1,118,533
#Source posts	379	960	379	960	379	960

S7.6 Moral Outrage Analysis

The estimation results for the analysis of moral outrage in replies are reported in Table S7.19. Additionally, the estimation results for moral outrage predicted by Digital Outrage Classifier (DOC) [312] and other-condemning emotions predicted by moral emotion classifier [365] are reported in Table S7.20. We find that the coefficient estimates of *Displayed* for moral outrage represented by the product of anger and disgust and for moral outrage predicted by DOC are quantitatively identical based on *SUEST* test ($\chi^2 = 0.00, p = 0.957$).

Table S7.19: Estimation results for moral outrage in replies across all misleading posts [Column (1)], political misleading posts [Column (2)], and non-political misleading posts [Column (3)]. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) All	(2) Politics	(3) Non-politics
Displayed	0.067*** (0.009)	0.062*** (0.013)	0.060*** (0.012)
HoursFromDisplay	0.000 (0.013)	−0.001 (0.022)	0.000 (0.008)
SourceAnger	0.033** (0.011)	0.030* (0.015)	0.025* (0.011)
SourceDisgust	0.027*** (0.007)	0.044* (0.017)	0.028*** (0.007)
SourceFear	0.001 (0.013)	−0.007 (0.022)	0.011 (0.013)
SourceJoy	−0.004 (0.009)	0.001 (0.008)	−0.016 (0.010)
SourceSadness	−0.002 (0.008)	0.007 (0.012)	−0.008 (0.009)
SourceSurprise	−0.037*** (0.010)	−0.030* (0.013)	−0.040*** (0.012)
PostAge	−0.002 (0.008)	−0.003 (0.012)	0.003 (0.008)
Intercept	−0.017 (0.009)	0.035* (0.014)	−0.062*** (0.010)
R^2	0.006	0.005	0.007
#Replies (N)	2,155,873	1,037,340	1,118,533
#Source posts	1339	379	960

	(1) Moral outrage (DOC)	(2) Other-condemning
Displayed	0.067*** (0.013)	0.132*** (0.015)
HoursFromDisplay	−0.004 (0.010)	−0.001 (0.017)
SourceAnger	0.055*** (0.012)	0.085*** (0.018)
SourceDisgust	0.017* (0.008)	0.004 (0.012)
SourceFear	0.015 (0.013)	0.011 (0.018)
SourceJoy	0.019 (0.025)	0.012 (0.028)
SourceSadness	0.008 (0.011)	0.015 (0.016)
SourceSurprise	−0.024 (0.014)	−0.045* (0.020)
PostAge	0.003 (0.007)	−0.006 (0.011)
Intercept	−0.017 (0.012)	−0.033 (0.017)
R^2	0.006	0.016
#Replies (N)	2,155,873	2,155,873
#Source posts	1339	1339

Table S7.20: Estimation results for moral outrage based on DOC [Column (1)] and other-condemning moral emotions [Column (2)] in replies. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

S7.7 Robustness Checks

S7.7.1 Variance Inflation Factors

The Variance Inflation Factors (VIFs) for the independent variables are reported in Table S7.21.

	(1) Sentiments	(2) Emotions
Displayed	1.11	1.12
HoursFromDisplay	1.51	1.52
SourcePositive	1.48	
SourceNegative	1.48	
SourceAnger		1.38
SourceDisgust		1.16
SourceFear		1.34
SourceJoy		1.33
SourceSadness		1.24
SourceSurprise		1.57
PostAge	1.43	1.44
Mean VIF	1.40	1.34

Table S7.21: VIFs for independent variables. Column (1) reports the VIFs for independent variables in regressions for sentiments in replies. Column (2) reports the VIFs for independent variables in regressions for emotions in replies.

S7.7.2 Test for Potential Autocorrelation

We test the correlations of sentiments and emotions in replies in chronological, shuffled, and reversed orders, respectively. The estimation results are reported in Table S7.22 and visualized in Fig. S7.4. Additionally, we incorporate lagged-dependent covariates, *i.e.*, average sentiments/emotions in previous direct replies, into the regression models. Given the high positive correlations between sentiments/emotions in replies and their corresponding ones in source posts (Table S7.23), we use their product as combined variables to mitigate potential multicollinearity issues. The results remain consistent and support our findings (Tables S7.24–S7.25).



Figure S7.4: The lagged correlation estimates across sentiments and emotions in replies in Chronological (C), Shuffled (S), and Reversed (R) orders, respectively. Shown are mean values with error bars representing 95% CIs. The full estimation results are reported in Table S7.22.

Table S7.22: The lagged coefficient estimates across sentiments and emotions in replies. $N = 1,686,262$. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) Sentiments in replies		(2) Emotions in replies					
	Positive	Negative	Anger	Disgust	Fear	Joy	Sadness	Surprise
Chronological	0.271*** (0.003)	0.324*** (0.003)	0.225*** (0.003)	0.201*** (0.003)	0.149*** (0.009)	0.182*** (0.003)	0.175*** (0.004)	0.184*** (0.003)
Shuffled	0.268*** (0.003)	0.323*** (0.002)	0.224*** (0.002)	0.200*** (0.003)	0.151*** (0.006)	0.180*** (0.004)	0.174*** (0.004)	0.184*** (0.003)
Reversed	0.268*** (0.006)	0.323*** (0.002)	0.224*** (0.002)	0.199*** (0.003)	0.150*** (0.005)	0.180*** (0.005)	0.174*** (0.004)	0.184*** (0.003)

S7.7.3 Shorter Bandwidth and Placebo Test

We repeat our analysis within a shorter bandwidth of 2 hours before and after the display of community notes. Specifically, the before-display period is 2 hours preceding the display of community notes, and the after-display period is from fifth to sixth hour (2 hours) after the note display. The initial four hours after the display of community notes are omitted to avoid potential cold-start contamination in the short bandwidth. The full estimation results are reported in Table S7.26 and Table S7.27. All the results, especially for anger and moral outrage, are robust and consistent with our main findings.

	<i>r</i> : Source – Lagged	
Positive	0.358***	
Negative	0.426***	
Anger	0.290***	
Disgust	0.228***	
Fear	0.342***	
Joy	0.246***	
Sadness	0.274***	
Surprise	0.230***	

	(1) Positive	(2) Negative
Displayed	−0.053*** (0.014)	0.117 *** (0.017)
HoursFromDisplay	0.004 (0.018)	−0.012 (0.027)
PriorPositiveMultiply	0.045*** (0.010)	−0.015 (0.011)
PriorNegativeMultiply	0.027 (0.019)	−0.078** (0.030)
PostAge	0.001 (0.010)	0.002 (0.015)
Intercept	−0.006 (0.015)	−0.001 (0.017)
<i>R</i> ²	0.007	0.010
#Replies (<i>N</i>)	2,154,534	2,154,534
#Source posts	1339	1339

Table S7.23: The Pearson correlations between source sentiments/emotions in original posts and lagged sentiments/emotions in replies with two-sided *p*-values. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table S7.24: Estimation results for positive [Column (1)] and negative [Column (2)] sentiments in replies. The product terms representing the combination of source sentiments and lagged sentiments in replies are included. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Additionally, we shift the cut-off point to 2 hours before the display of community notes and conduct placebo test based on replies within 2 hours before and after the shifted cut-off point. The full estimation results are reported in Table S7.28 and Table S7.29. All the RDD estimators, *i.e.*, the coefficient estimates of *Displayed* across sentiments and emotions in replies, are statistically not significant. This indicates that the changes in sentiments and emotions are specific to the display of community notes.

S7.7.4 Analysis Without Limiting the Number of Replies

In the main analysis, we restrict our analysis to posts that receive at least 5 replies both before and after the display of community notes. This ensures that each post has sufficient replies both before and after note display for a meaningful comparison. To test the robustness of our findings, we repeat our analysis without this restriction. The estimation results for sentiments and emotions in replies are reported in Table S7.30 and Table S7.31. All the results remain consistent with our main findings.

S7.7.5 Month-Year Fixed Effects

We consider the month-year fixed effects at two levels: post level and reply level. At the post level, we incorporate month-year fixed effects based on the creation time of misleading posts. The estimation results

Table S7.25: Estimation results for anger [Column (1)], disgust [Column (2)], fear [Column (3)], joy [Column (4)], sadness [Column (5)], surprise [Column (6)], and moral outrage [Column (7)] in replies. The product terms representing the combination of source emotions and lagged emotions in replies are included. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) Anger	(2) Disgust	(3) Fear	(4) Joy	(5) Sadness	(6) Surprise	(7) Moral outrage
Displayed	0.089*** (0.014)	0.031*** (0.009)	-0.003 (0.007)	-0.008 (0.009)	-0.003 (0.012)	-0.052*** (0.012)	0.074*** (0.009)
HoursFromDisplay	-0.001 (0.018)	-0.004 (0.009)	-0.003 (0.004)	-0.006 (0.007)	-0.007 (0.006)	-0.009 (0.012)	0.002 (0.012)
PriorAngerMultiply	0.038* (0.017)	0.008 (0.007)	0.003 (0.005)	-0.017** (0.006)	-0.013 (0.009)	-0.032*** (0.009)	0.020 (0.012)
PriorDisgustMultiply	-0.002 (0.004)	0.031*** (0.005)	-0.005** (0.002)	-0.004** (0.002)	-0.005*** (0.001)	-0.001 (0.003)	0.009** (0.003)
PriorFearMultiply	-0.006 (0.005)	-0.004 (0.005)	0.047*** (0.008)	-0.008** (0.003)	-0.004 (0.003)	-0.005 (0.004)	-0.005 (0.004)
PriorJoyMultiply	-0.011** (0.004)	-0.004 (0.003)	0.000 (0.004)	0.025** (0.009)	0.002 (0.002)	0.001 (0.004)	-0.006* (0.003)
PriorSadnessMultiply	-0.011 (0.006)	-0.009 (0.005)	0.001 (0.004)	0.005 (0.007)	0.029*** (0.007)	0.003 (0.006)	-0.009 (0.005)
PriorSurpriseMultiply	-0.022* (0.011)	-0.017 (0.009)	-0.007 (0.005)	-0.003 (0.005)	0.015 (0.017)	0.040** (0.014)	-0.018 (0.009)
PostAge	-0.005 (0.010)	0.002 (0.007)	0.004 (0.004)	0.007 (0.005)	0.006 (0.004)	0.013 (0.008)	-0.003 (0.008)
Intercept	-0.021 (0.015)	-0.010 (0.011)	-0.013 (0.007)	0.009 (0.011)	-0.003 (0.012)	0.011 (0.013)	-0.018 (0.011)
R^2	0.004	0.005	0.006	0.002	0.002	0.004	0.002
#Replies (N)	2,154,534	2,154,534	2,154,534	2,154,534	2,154,534	2,154,534	2,154,534
#Source posts	1339	1339	1339	1339	1339	1339	1339

across sentiments and emotions are reported in Table S7.32 and Table S7.33. The estimation results for moral outrage are reported in Column (1) of Table S7.36. At the reply level, we incorporate month-year fixed effects based on the creation time of replies. The estimation results across sentiments and emotions are reported in Table S7.34 and Table S7.35. The estimation results for moral outrage are reported in Column (2) of Table S7.36. All the results are robust and consistent with our main findings.

S7.7.6 Alternative Lexicon-Based Sentiments and Emotions

We use VADER sentiment lexicon [366] and NRC emotion lexicon [163] as alternative methods for measuring sentiments and emotions. The estimation results across sentiments and emotions based on the alternative approaches are reported in Table S7.37 and Table S7.38. Additionally, the estimation results for moral outrage based on NRC emotion lexicon are reported in Column (3) of Table S7.36. All the results are robust and consistent with our main findings.

S7.7.7 Expansion With Sentiments/Emotions in Community Notes

We expand our models by incorporating sentiment (emotions) in community notes and their interactions with *Displayed*. The estimation results across sentiments and emotions are reported in Table S7.39 and Table S7.40. Additionally, the estimation results for moral outrage are reported in Column (4) of Table S7.36. All the results are robust and consistent with our main findings.

	(1) Positive	(2) Negative
Displayed	−0.106 (0.056)	0.152* (0.063)
HoursFromDisplay	0.852 (0.512)	−0.792 (0.539)
SourcePositive	0.075 (0.045)	−0.026 (0.038)
SourceNegative	−0.020 (0.029)	0.071* (0.032)
PostAge	−0.004 (0.025)	−0.016 (0.032)
Intercept	−0.153 (0.094)	0.130 (0.103)
R^2	0.010	0.010
#Replies (N)	278,754	278,754
#Source posts	1280	1280

Table S7.26: Estimation results for positive sentiment [Column (1)] and negative sentiment [Column (2)] in replies within 2-hour bandwidth. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

S7.7.8 Expansion With Helpfulness Scores of Community Notes

Misleading posts that garner significant public concern are often subject to external factors such as heightened public criticism [50, 67, 76]. These factors may contribute to the display of community notes with high helpfulness scores and increased negativity in replies, thereby introducing potential confounds. In addition to displaying helpful community notes on the corresponding misleading posts directly, Community Notes official X account (@HelpfulNotes) automatically quote community notes and associated misleading posts to promote the helpful context based on certain criteria, for example, having a high helpfulness score. This could influence how users interact with the original misleading posts and complicate the effect of community notes display. To mitigate these concerns, we repeat our analysis by including helpfulness scores of community notes. We first calculate community notes' helpfulness scores based on the same algorithm used in production (the open-source algorithm is available at <https://github.com/twitter/communitynotes/tree/main>). Subsequently, we incorporate the variable *NoteScore* indicating helpfulness scores of community notes and its interaction with *Displayed* in the RDD models. The estimation results for sentiments and emotions in replies are reported in Table S7.41 and Table S7.42. We find that the coefficient estimates of *Displayed* $\times$ *NoteScore* are not statistically significant across sentiments and emotions in replies (each $p > 0.05$). This suggests that quoting community notes by Community Notes official X account has no significant effect on the sentiments and emotions in replies received by misleading posts.

S7.7.9 Robustness to the Development of Community Notes Program

We incorporate a continuous variable, *MFRO* (*i.e.*, Months From the Roll-Out for source posts or replies), to investigate how the effect of community notes changes over time as the program develops. Table S7.43 and Table S7.44 report estimates results with *MFRO* based on post

Table S7.27: Estimation results for anger [Column (1)], disgust [Column (2)], fear [Column (3)], joy [Column (4)], sadness [Column (5)], surprise [Column (6)], and moral outrage [Column (7)] in replies within 2-hour bandwidth. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) Anger	(2) Disgust	(3) Fear	(4) Joy	(5) Sadness	(6) Surprise	(7) Moral outrage
Displayed	0.096* (0.038)	0.019 (0.025)	−0.002 (0.021)	−0.057 (0.060)	0.008 (0.020)	−0.093*** (0.025)	0.073** (0.028)
HoursFromDisplay	−0.389 (0.371)	−0.055 (0.238)	0.038 (0.219)	0.607 (0.553)	−0.029 (0.222)	0.396 (0.255)	−0.329 (0.305)
SourceAnger	0.059*** (0.016)	0.005 (0.014)	0.019* (0.008)	−0.008 (0.010)	0.027* (0.011)	−0.021* (0.010)	0.016 (0.013)
SourceDisgust	−0.007 (0.012)	0.062*** (0.012)	−0.007 (0.005)	−0.002 (0.009)	−0.011* (0.005)	0.007 (0.009)	0.023* (0.011)
SourceFear	0.013 (0.018)	−0.008 (0.013)	0.064*** (0.012)	−0.013 (0.009)	0.019* (0.009)	0.007 (0.010)	−0.007 (0.016)
SourceJoy	0.006 (0.022)	−0.011 (0.010)	0.001 (0.008)	0.049 (0.034)	0.026*** (0.007)	−0.004 (0.010)	−0.012 (0.012)
SourceSadness	0.004 (0.015)	−0.014 (0.010)	0.007 (0.008)	0.013 (0.014)	0.048*** (0.010)	−0.004 (0.011)	−0.010 (0.013)
SourceSurprise	−0.056** (0.017)	−0.027* (0.013)	−0.006 (0.009)	0.003 (0.014)	0.029* (0.015)	0.081*** (0.017)	−0.042** (0.014)
PostAge	−0.005 (0.018)	0.022 (0.017)	0.008 (0.013)	0.001 (0.015)	−0.001 (0.013)	−0.004 (0.013)	0.020 (0.016)
Intercept	0.065 (0.070)	0.012 (0.043)	−0.018 (0.040)	−0.120 (0.098)	−0.015 (0.043)	−0.061 (0.047)	0.068 (0.057)
R^2	0.009	0.008	0.004	0.005	0.003	0.008	0.004
#Replies (N)	278,754	278,754	278,754	278,754	278,754	278,754	278,754
#Source posts	1280	1280	1280	1280	1280	1280	1280

creation. Table S7.45 and Table S7.46 report estimates results with *MFRO* based on reply creation. All coefficient estimates of *Displayed* $\times$ *MFRO* are not statistically significant (each $p > 0.05$). This suggests that the effect of community notes is stable relative to the development of Community Notes program.

	(1) Positive	(2) Negative
Displayed	0.043 (0.028)	-0.039 (0.028)
HoursFromDisplay	-0.012 (0.011)	0.009 (0.010)
SourcePositive	0.064 (0.044)	-0.011 (0.037)
SourceNegative	-0.026 (0.028)	0.080* (0.032)
PostAge	-0.001 (0.010)	-0.005 (0.013)
Intercept	-0.017 (0.026)	0.015 (0.033)
R^2	0.007	0.008
#Replies (N)	373,248	373,248
#Source posts	1282	1282

Table S7.28: Estimation results for positive sentiment [Column (1)] and negative sentiment [Column (2)] in replies within 2-hour bandwidth. The cut-off point is shifted to 2 hours before the display of community notes for placebo test. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table S7.29: Estimation results for anger [Column (1)], disgust [Column (2)], fear [Column (3)], joy [Column (4)], sadness [Column (5)], surprise [Column (6)], and moral outrage [Column (7)] in replies within 2-hour bandwidth. The cut-off point is shifted to 2 hours before the display of community notes for placebo test. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

[illegible]

	(1) Positive	(2) Negative
Displayed	−0.055*** (0.016)	0.114*** (0.020)
HoursFromDisplay	−0.003 (0.016)	0.000 (0.020)
SourcePositive	0.040* (0.018)	0.014 (0.021)
SourceNegative	−0.048** (0.016)	0.113*** (0.023)
PostAge	0.007 (0.009)	−0.006 (0.012)
Post: MonthYear Intercept	✓ −0.017 (0.018)	✓ −0.045 (0.037)
R^2	0.008	0.014
#Replies (N)	2,155,873	2,155,873
#Source posts	1339	1339

Table S7.32: Estimation results for positive sentiment [Column (1)] and negative sentiment [Column (2)] in replies. Post-level month-year fixed effects are included. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table S7.33: Estimation results for emotions in replies: anger [Column (1)], disgust [Column (2)], fear [Column (3)], joy [Column (4)], sadness [Column (5)], and surprise [Column (6)]. Post-level month-year fixed effects are included. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) Anger	(2) Disgust	(3) Fear	(4) Joy	(5) Sadness	(6) Surprise
Displayed	0.077*** (0.010)	0.023* (0.009)	−0.006 (0.007)	−0.015 (0.010)	−0.006 (0.011)	−0.043*** (0.009)
HoursFromDisplay	−0.004 (0.017)	−0.007 (0.010)	−0.001 (0.005)	−0.005 (0.011)	−0.002 (0.008)	−0.005 (0.010)
SourceAnger	0.076*** (0.012)	0.008 (0.011)	0.019* (0.008)	−0.008 (0.008)	0.020 (0.013)	−0.037*** (0.010)
SourceDisgust	−0.003 (0.009)	0.071*** (0.015)	−0.002 (0.004)	−0.010 (0.006)	−0.011 (0.006)	−0.003 (0.007)
SourceFear	0.009 (0.014)	0.008 (0.011)	0.068*** (0.011)	−0.019** (0.007)	0.007 (0.009)	0.004 (0.012)
SourceJoy	0.004 (0.015)	0.000 (0.008)	0.007 (0.006)	0.032 (0.017)	0.026*** (0.007)	0.000 (0.011)
SourceSadness	0.004 (0.010)	−0.003 (0.009)	0.009 (0.006)	0.008 (0.009)	0.052*** (0.008)	−0.008 (0.009)
SourceSurprise	−0.059*** (0.011)	−0.010 (0.012)	0.002 (0.007)	−0.001 (0.010)	0.016 (0.013)	0.067*** (0.013)
PostAge	−0.002 (0.010)	0.004 (0.008)	0.004 (0.004)	0.008 (0.007)	0.006 (0.005)	0.010 (0.006)
Post: MonthYear Intercept	✓ −0.047* (0.020)	✓ 0.015 (0.021)	✓ 0.032 (0.022)	✓ −0.022 (0.012)	✓ −0.006 (0.020)	✓ 0.001 (0.017)
R^2	0.015	0.006	0.005	0.003	0.003	0.009
#Replies (N)	2,155,873	2,155,873	2,155,873	2,155,873	2,155,873	2,155,873
#Source posts	1339	1339	1339	1339	1339	1339

Table S7.34: Estimation results for positive sentiment [Column (1)] and negative sentiment [Column (2)] in replies. Reply-level month-year fixed effects are included. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) Positive	(2) Negative
Displayed	−0.057** (0.018)	0.115 *** (0.021)
HoursFromDisplay	−0.002 (0.016)	−0.001 (0.020)
SourcePositive	0.039* (0.018)	0.014 (0.022)
SourceNegative	−0.049** (0.017)	0.113 *** (0.023)
PostAge	0.003 (0.009)	−0.002 (0.012)
Reply: MonthYear	✓	✓
Intercept	−0.017 (0.017)	−0.050 (0.037)
R^2	0.008	0.014
#Replies (N)	2,155,873	2,155,873
#Source posts	1339	1339

Table S7.35: Estimation results for emotions in replies: anger [Column (1)], disgust [Column (2)], fear [Column (3)], joy [Column (4)], sadness [Column (5)], and surprise [Column (6)]. Reply-level month-year fixed effects are included. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) Anger	(2) Disgust	(3) Fear	(4) Joy	(5) Sadness	(6) Surprise
Displayed	0.078*** (0.011)	0.023* (0.009)	−0.007 (0.007)	−0.016 (0.010)	−0.007 (0.011)	−0.043*** (0.009)
HoursFromDisplay	−0.005 (0.017)	−0.008 (0.010)	−0.001 (0.005)	−0.003 (0.012)	−0.002 (0.008)	−0.004 (0.010)
SourceAnger	0.076*** (0.012)	0.008 (0.011)	0.019* (0.008)	−0.008 (0.008)	0.020 (0.013)	−0.037*** (0.010)
SourceDisgust	−0.002 (0.008)	0.072*** (0.015)	−0.002 (0.004)	−0.012* (0.006)	−0.012* (0.006)	−0.004 (0.007)
SourceFear	0.009 (0.014)	0.009 (0.011)	0.067*** (0.011)	−0.020** (0.007)	0.007 (0.009)	0.003 (0.012)
SourceJoy	0.004 (0.015)	0.001 (0.008)	0.007 (0.006)	0.030 (0.018)	0.024*** (0.007)	0.000 (0.011)
SourceSadness	0.004 (0.010)	−0.002 (0.009)	0.008 (0.006)	0.006 (0.009)	0.052*** (0.008)	−0.008 (0.008)
SourceSurprise	−0.060*** (0.011)	−0.009 (0.012)	0.002 (0.007)	−0.002 (0.010)	0.016 (0.013)	0.067*** (0.012)
PostAge	−0.001 (0.009)	0.003 (0.008)	0.005 (0.004)	0.005 (0.007)	0.006 (0.004)	0.011 (0.006)
Reply: MonthYear	✓	✓	✓	✓	✓	✓
Intercept	−0.050* (0.020)	0.016 (0.021)	0.028 (0.022)	−0.024 (0.012)	−0.008 (0.019)	0.002 (0.017)
R^2	0.015	0.006	0.005	0.003	0.003	0.009
#Replies (N)	2,155,873	2,155,873	2,155,873	2,155,873	2,155,873	2,155,873
#Source posts	1339	1339	1339	1339	1339	1339

Table S7.36: Estimation results for moral outrage across robustness checks: post-level month-year fixed effects [Column (1)], reply-level month-year fixed effects [Column (2)], alternative sentiment and emotion lexicons [Column (3)], and expansion with emotions in community notes [Column (4)]. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) Post: month-year	(2) Reply: month-year	(3) Alternative lexicons	(4) Emotions in community notes
Displayed	0.066*** (0.008)	0.066*** (0.009)	0.054*** (0.009)	0.068*** (0.008)
HoursFromDisplay	−0.001 (0.013)	−0.001 (0.013)	−0.021 (0.016)	−0.001 (0.013)
NoteAnger				0.001 (0.011)
NoteDisgust				0.002 (0.008)
NoteFear				0.002 (0.009)
NoteJoy				0.006 (0.011)
NoteSadness				−0.008 (0.009)
NoteSurprise				−0.023* (0.010)
Displayed × NoteAnger				0.021 (0.012)
Displayed × NoteDisgust				−0.010 (0.008)
Displayed × NoteFear				0.004 (0.011)
Displayed × NoteJoy				−0.007 (0.010)
Displayed × NoteSadness				0.004 (0.009)
Displayed × NoteSurprise				0.008 (0.011)
SourceAnger	0.032** (0.011)	0.032** (0.011)	0.032* (0.013)	0.030** (0.011)
SourceDisgust	0.025*** (0.007)	0.026*** (0.008)	0.025** (0.008)	0.024** (0.007)
SourceFear	−0.002 (0.012)	−0.001 (0.012)	−0.005 (0.007)	0.000 (0.012)
SourceJoy	−0.008 (0.008)	−0.007 (0.008)	0.009 (0.009)	−0.007 (0.009)
SourceSadness	−0.005 (0.008)	−0.005 (0.008)	0.022 (0.016)	−0.001 (0.009)
SourceSurprise	−0.038*** (0.010)	−0.038*** (0.010)	−0.004 (0.005)	−0.035*** (0.010)
PostAge	−0.001 (0.008)	−0.001 (0.008)	0.016 (0.010)	−0.001 (0.008)
Intercept	−0.027 (0.018)	−0.026 (0.018)	−0.013 (0.008)	−0.017 (0.009)
R^2	0.006	0.006	0.003	0.006
#Replies (N)	2,155,873	2,155,873	2,155,873	2,155,873
#Source posts	1339	1339	1339	1339

Table S7.37: Estimation results for sentiment polarity and emotions in replies: polarity [Column (1)], anger [Column (2)], disgust [Column (3)], and fear [Column (4)]. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) Polarity	(2) Anger	(3) Disgust	(4) Fear
Displayed	−0.044** (0.015)	0.035** (0.014)	0.062*** (0.011)	−0.006 (0.010)
HoursFromDisplay	0.001 (0.012)	−0.015 (0.012)	−0.029 (0.021)	0.002 (0.006)
SourcePolarity	0.072*** (0.016)			
SourceAnger		0.058*** (0.012)	0.015 (0.018)	0.026* (0.013)
SourceDisgust		0.004 (0.007)	0.069*** (0.018)	−0.004 (0.007)
SourceFear		0.022** (0.008)	−0.018 (0.010)	0.099*** (0.010)
SourceJoy		0.004 (0.009)	0.004 (0.011)	−0.013 (0.008)
SourceSadness		0.010 (0.012)	0.006 (0.016)	−0.017** (0.006)
SourceSurprise		−0.011 (0.006)	−0.006 (0.006)	−0.008 (0.006)
PostAge	−0.001 (0.007)	0.009 (0.007)	0.023 (0.015)	0.000 (0.005)
Intercept	0.011 (0.012)	−0.009 (0.010)	−0.016 (0.012)	0.001 (0.010)
R^2	0.006	0.005	0.007	0.013
#Replies (N)	2,155,873	2,155,873	2,155,873	2,155,873
#Source posts	1339	1339	1339	1339

Table S7.38: Estimation results for emotions in replies: joy [Column (1)], sadness [Column (2)], and surprise [Column (3)]. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) Joy	(2) Sadness	(3) Surprise
Displayed	0.007 (0.012)	0.010 (0.016)	−0.007 (0.013)
HoursFromDisplay	0.002 (0.007)	0.004 (0.006)	0.000 (0.006)
SourceAnger	−0.005 (0.006)	0.003 (0.011)	−0.006 (0.007)
SourceDisgust	0.002 (0.006)	0.009 (0.009)	−0.005 (0.009)
SourceFear	−0.011 (0.006)	−0.014 (0.009)	−0.013 (0.007)
SourceJoy	0.074*** (0.009)	0.010 (0.013)	−0.009 (0.008)
SourceSadness	0.020** (0.007)	0.056** (0.017)	0.003 (0.010)
SourceSurprise	0.006 (0.006)	−0.016* (0.006)	0.047*** (0.010)
PostAge	0.000 (0.005)	−0.004 (0.005)	−0.005 (0.005)
Intercept	−0.002 (0.008)	−0.002 (0.014)	0.002 (0.010)
R^2	0.006	0.004	0.003
#Replies (N)	2,155,873	2,155,873	2,155,873
#Source posts	1339	1339	1339

	(1) Positive	(2) Negative
Displayed	−0.054** (0.017)	0.113 *** (0.022)
HoursFromDisplay	−0.003 (0.013)	−0.002 (0.023)
NotePositive	−0.002 (0.011)	0.036* (0.016)
NoteNegative	0.023 (0.013)	0.031 (0.020)
Displayed × NotePositive	0.006 (0.010)	−0.003 (0.013)
Displayed × NoteNegative	−0.024 (0.013)	0.015 (0.019)
SourcePositive	0.040* (0.019)	0.017 (0.021)
SourceNegative	−0.049** (0.017)	0.110 *** (0.023)
PostAge	0.005 (0.008)	−0.005 (0.014)
Intercept	0.013 (0.013)	−0.028 (0.018)
R^2	0.007	0.015
#Replies (N)	2,155,873	2,155,873
#Source posts	1339	1339

Table S7.39: Estimation results for positive sentiment [Column (1)] and negative sentiment [Column (2)] in replies. Sentiments in community notes and their interactions with *Displayed* are included. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table S7.40: Estimation results for anger [Column (1)], disgust [Column (2)], fear [Column (3)], joy [Column (4)], sadness [Column (5)], and surprise [Column (6)] in replies. Emotions in community notes and their interactions with *Displayed* are included. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	(1) Anger	(2) Disgust	(3) Fear	(4) Joy	(5) Sadness	(6) Surprise
Displayed	0.078*** (0.010)	0.023* (0.010)	−0.005 (0.007)	−0.016 (0.009)	−0.010 (0.010)	−0.043*** (0.008)
HoursFromDisplay	−0.006 (0.018)	−0.007 (0.009)	−0.003 (0.004)	−0.003 (0.010)	0.001 (0.007)	−0.004 (0.009)
NoteAnger	0.034** (0.011)	−0.014 (0.013)	−0.017 (0.009)	0.026* (0.011)	0.010 (0.012)	−0.020* (0.009)
NoteDisgust	−0.005 (0.007)	0.002 (0.009)	−0.004 (0.004)	0.015 (0.009)	−0.009 (0.008)	0.009 (0.008)
NoteFear	0.017 (0.011)	−0.012 (0.012)	0.042* (0.018)	−0.004 (0.011)	−0.007 (0.010)	−0.021* (0.010)
NoteJoy	0.019 (0.012)	0.009 (0.014)	−0.012* (0.006)	0.006 (0.008)	0.002 (0.009)	−0.005 (0.011)
NoteSadness	−0.002 (0.010)	−0.008 (0.013)	−0.012* (0.006)	0.016 (0.009)	0.059*** (0.012)	−0.004 (0.011)
NoteSurprise	−0.033** (0.012)	−0.020 (0.011)	−0.001 (0.006)	0.004 (0.007)	−0.008 (0.010)	0.026** (0.009)
Displayed × NoteAnger	0.006 (0.009)	0.031* (0.015)	0.004 (0.007)	−0.026** (0.010)	−0.009 (0.011)	0.001 (0.008)
Displayed × NoteDisgust	0.001 (0.007)	−0.011 (0.009)	0.005 (0.003)	−0.001 (0.007)	0.003 (0.007)	−0.003 (0.007)
Displayed × NoteFear	0.007 (0.012)	0.006 (0.012)	−0.003 (0.015)	−0.020 (0.011)	−0.002 (0.009)	0.012 (0.011)
Displayed × NoteJoy	0.000 (0.011)	0.001 (0.011)	0.007 (0.006)	−0.006 (0.012)	0.004 (0.008)	−0.007 (0.010)
Displayed × NoteSadness	0.007 (0.009)	0.003 (0.012)	0.005 (0.005)	−0.026** (0.008)	0.003 (0.012)	0.010 (0.010)
Displayed × NoteSurprise	0.010 (0.012)	0.010 (0.012)	0.009 (0.007)	0.005 (0.010)	−0.003 (0.011)	−0.007 (0.010)
SourceAnger	0.066*** (0.013)	0.008 (0.011)	0.015* (0.007)	−0.009 (0.008)	0.013 (0.011)	−0.027** (0.010)
SourceDisgust	−0.004 (0.009)	0.067*** (0.014)	−0.003 (0.005)	−0.013* (0.006)	−0.008 (0.006)	0.000 (0.008)
SourceFear	0.008 (0.012)	0.012 (0.011)	0.054*** (0.008)	−0.019* (0.008)	0.006 (0.008)	0.010 (0.011)
SourceJoy	0.003 (0.013)	−0.002 (0.008)	0.004 (0.007)	0.032 (0.018)	0.026*** (0.007)	0.003 (0.010)
SourceSadness	0.010 (0.010)	−0.001 (0.010)	0.005 (0.006)	0.005 (0.008)	0.043*** (0.009)	−0.007 (0.009)
SourceSurprise	−0.052*** (0.011)	−0.009 (0.012)	−0.008 (0.007)	0.002 (0.010)	0.017 (0.011)	0.068*** (0.013)
PostAge	−0.001 (0.010)	0.005 (0.008)	0.006 (0.004)	0.006 (0.006)	0.003 (0.004)	0.009 (0.006)
Intercept	−0.019 (0.011)	−0.005 (0.010)	0.001 (0.007)	0.004 (0.009)	0.002 (0.011)	0.011 (0.009)
R^2	0.018	0.006	0.007	0.003	0.007	0.010
#Replies (N)	2,155,873	2,155,873	2,155,873	2,155,873	2,155,873	2,155,873
#Source posts	1339	1339	1339	1339	1339	1339

	(1) Positive	(2) Negative
Displayed	−0.053*** (0.015)	0.110 *** (0.019)
HoursFromDisplay	−0.003 (0.014)	0.001 (0.020)
MFRO _{reply}	0.006 (0.017)	−0.022 (0.020)
Displayed × MFRO _{reply}	−0.017 (0.019)	0.044 (0.023)
SourcePositive	0.039* (0.019)	0.016 (0.021)
SourceNegative	−0.048** (0.017)	0.112 *** (0.023)
PostAge	0.006 (0.009)	−0.007 (0.013)
Intercept	0.014 (0.014)	−0.030 (0.018)
R^2	0.007	0.014
#Replies (N)	2,155,873	2,155,873
#Source posts	1339	1339

Table S7.45: Estimation results for positive [Column (1)] and negative [Column (2)] sentiments in replies. The control variable *MFRO_{reply}* and its interaction with *Displayed* is included. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table S7.46: Estimation results for anger [Column (1)], disgust [Column (2)], fear [Column (3)], joy [Column (4)], sadness [Column (5)], surprise [Column (6)], and moral outrage [Column (7)] in replies. The control variable *MFR0_{reply}* and its interaction with *Displayed* is included. Reported are coefficient estimates with post-clustered standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

[illegible]

S8

Supplementary Materials for Chapter 8

S8.1 Status Updates on Community Notes

When a community note is first submitted, it is assigned the status of “Needs More Ratings” (see Fig. 8.1 in the main paper). When community notes receive enough ratings from raters with a broad range of perspectives, their statuses will be updated to either “Currently Rated Helpful” or “Currently Rated Not Helpful” by the note selection algorithm [387]. Then, the note selection algorithm updates the note status regularly by incorporating newly submitted ratings. Therefore, community notes that have received helpful status and been displayed on the misleading posts can return to “Needs More Ratings” status or change to “Currently Rated Not Helpful,” thereby losing their display status.

S8.2 Reproduction of Note Selection Algorithm

The distribution of estimated Core note intercepts (helpfulness) and estimated Core note factors (polarization) from the note selection algorithm is shown in Fig. S8.1. To ensure the reliability of our reproduction, we compare the note status generated by the reproduced algorithm with the actual production status. Each community note can receive one of the three statuses: Currently Rated Helpful, Needs More Ratings, and Currently Rated Not Helpful. Overall, 99.8% of notes in our replication match their corresponding production statuses, highlighting the reliability of our reproduction.

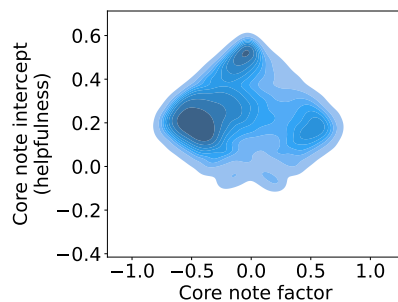


Figure S8.1: The distribution of estimated Core note intercepts (helpfulness) and estimated Core note factors (polarization) from the note selection algorithm.

S8.3 Matrix Factorization for Rater Similarity

The visualization of MSE training losses of adapted MF algorithm, its comparison with original note selection algorithm, and the distributions of rater similarity across different groups are shown in Fig. S8.2.

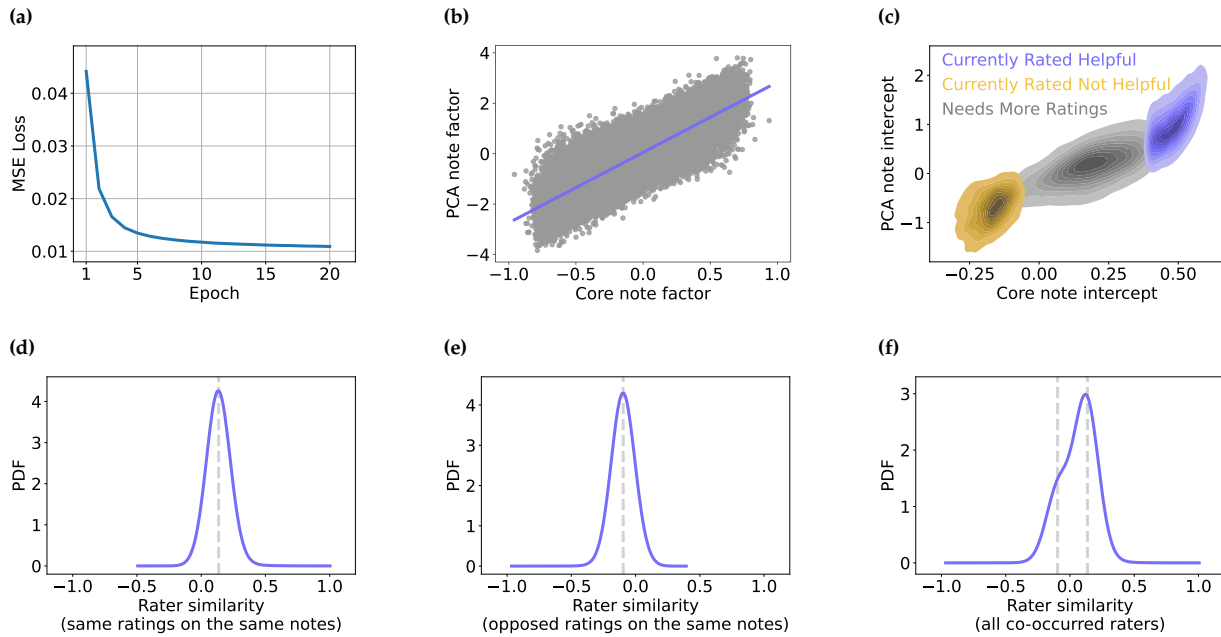


Figure S8.2: Analysis of MF-based model for rater representations. (a) The MSE losses over 20 epochs of training. (b) The distribution of estimated Core note factors from the note selection algorithm and estimated PCA note factors from the MF-based model for rater representations. (c) The distribution of estimated Core note intercepts from the note selection algorithm and estimated note intercepts from the MF-based model for rater representations. (d) The distribution of rater similarity for the same ratings on the same notes. (e) The distribution of rater similarity for the opposed ratings on the same notes. (f) The overall distribution of rater similarity across all co-rating pairs.

S8.4 Characteristics of Posts and Post Authors

We consider various characteristics of posts and their authors, which could influence both the spread of misleading posts and the evaluation of associated community notes.

Post characteristics.

- ▶ *Topics.* The dataset includes four topic categories—*i.e.*, Politics, Science, Health, and Economy. Each post may be associated with multiple topics, and each topic category is represented as a binary variable.
- ▶ *Sentiments.* Each post is annotated with positive and negative sentiment scores, represented as continuous variables that capture the polarity of emotional expression in the content.
- ▶ *Media.* A binary variable indicating whether the post contain media elements (= 1) or not (= 0).

Post author characteristics.

- ▶ *Verified.* A binary variable indicating whether a given post account is verified (= 1) or not (= 0).
- ▶ *Account age.* A continuous variable indicating the account age (in days) of a given post account since its creation.
- ▶ *Followers.* A count variable indicating the number of followers of a give post account.
- ▶ *Followees.* A count variable indicating the number of followees of a give post account.
- ▶ *Misinformation exposure score and partisan score.* Following the approach proposed by Mosleh and Rand [17], the misinformation exposure score of an author, ranging from 0 to 1, represents the proportion of followed public figures whose statements were rated as false by PolitiFact. The

partisan score, ranging from -1 (Democrat) to 1 (Republican), captures the ideological leaning of an author based on the political orientation of followed public figures.

S8.5 Sensitivity Analysis

The estimation results for our sensitivity analysis are shown in Fig. S8.3. Fig. S8.3a presents estimated changes in rating leaning for all displayed notes after initial display, across different post content and post author characteristics. Fig. S8.3b presents estimated changes in rating leaning for similar and dissimilar raters, across different post content and post author characteristics.

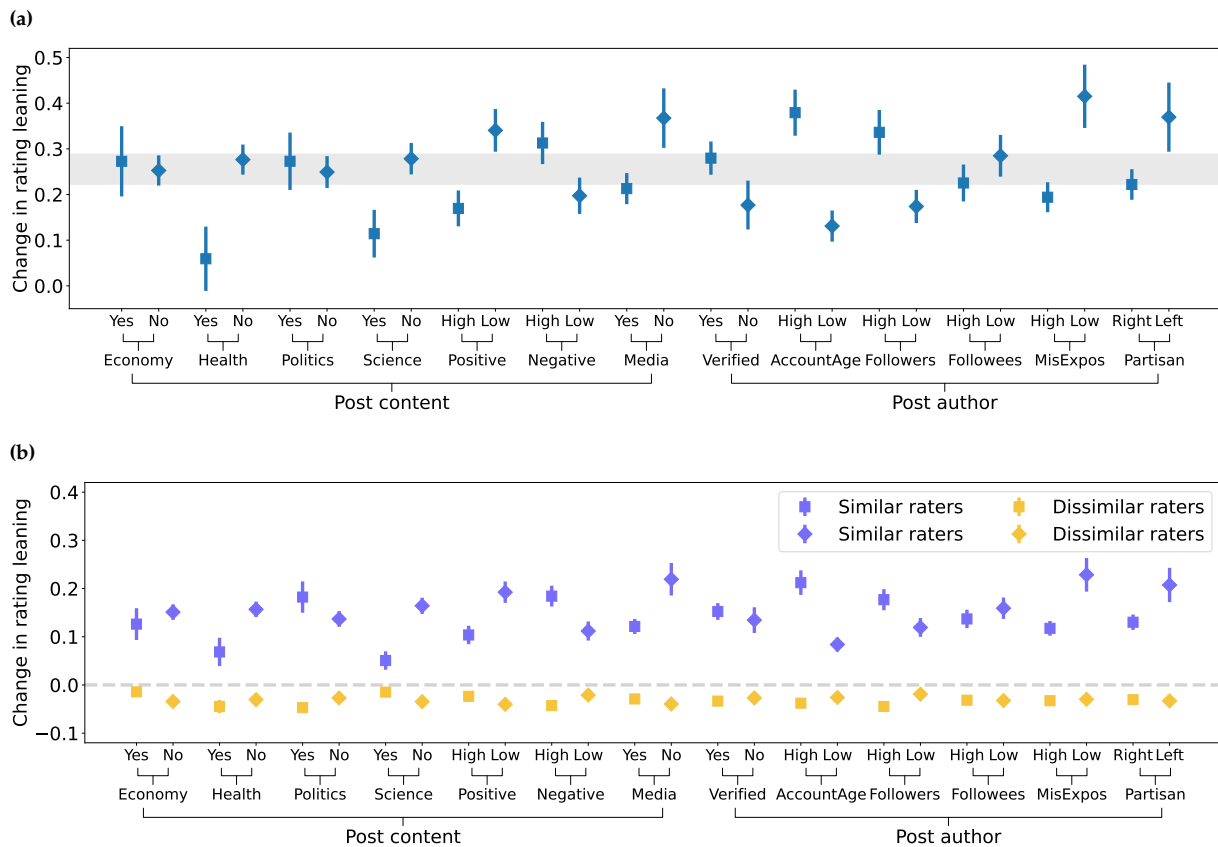


Figure S8.3: Sensitivity analysis. (a) Estimated changes in rating leaning for all displayed notes after initial display, across different post content and post author characteristics. The error bars represent 95% CIs. (b) Estimated changes in rating leaning for similar and dissimilar raters, across different post content and post author characteristics. The error bars represent 95% CIs.

S8.6 Evaluation of Domain Bias and Quality

We assess the political bias of external domains in community notes using the media bias dataset from Media Bias/Fact Check—the most comprehensive media bias resource on the internet (<https://mediabiasfactcheck.com>). Specifically, the domains in the categories of “Least Biased” or “Pro-Science” are coded as unbiased ($= 0$). The domains in the categories of “Left” ($= -1$) or “Left-Center” ($= -0.5$) are coded as left-leaning. The domains the categories of “Right” ($= 1$) or “Right-Center” ($= 0.5$)

are coded as right-leaning. When a note cites multiple domains, we average their bias scores, with positive values indicating right-leaning and negative values indicating left-leaning. Additionally, we measure the information quality of domains in community notes based on a large-scale domain-quality dataset compiled by Lin et al. [411]. The rating score ranges from 0 to 1, with a value of 0.5 or higher indicating high quality. We average the ratings for notes containing multiple domains.

S9

Supplementary Materials for Chapter 9

S9.1 Misleading Criteria

The criteria for labeling misleading posts in X's Community Notes system is shown in Fig. S9.1. The listed criteria suggest that the Community Notes system aims to provide corrective or contextual information for not only factually incorrect content or alter images and videos, but also posts that present accurate information in a misleading manner, such as, through omission of context, exaggeration, or distortion of the original meaning. This is consistent with the definition of misinformation in our study. Therefore, we consider misleading posts annotated by community notes on X to be manifestations of misinformation.

×

Add a note

Why do you believe this post may be misleading?

It contains a factual error

It contains a digitally altered photo or video

It contains outdated information that may be misleading

It is a misrepresentation or missing important context

It presents an unverified claim as a fact

It is a joke or satire that might be misinterpreted as a fact

Other

☐

☐

☐

☐

☐

☐

☐

Write a note with context that you believe should be shown with the post to keep others informed.

[See examples](#)

Your explanation

Be precise — providing links to outside sources is required.

Did you link to sources you believe most people would consider trustworthy?

Yes

No

☐

☐

Figure S9.1: The note-writing interface for Community Notes contributors, showing the specific criteria for labeling posts as misleading. Contributors are also required to provide explanations and cite external sources to justify their assessments.

S9.2 GPT Prompt

The full GPT-4.1 prompt used in our analysis is provided below. In addition to the definition of misinformation, we adopt the definitions of claim and opinion in a previous study [407].

```

1 You are an assistant that evaluates tweet content for two
  independent dimensions: content type and misleadingness.
2
3 ## Dimension 1: Content Type
4
5 Assess whether the tweet expresses a factual/verifiable claim and/
  or a subjective opinion.
6
7 ### Claim
8
9 A factual claim is a statement that explicitly presents some
  verifiable facts.
10 Statements with subjective components like opinions can also be
    factual claims if they explicitly present objectively
    verifiable facts.
11
12 Indicators that a tweet contains a factual claim:
13 - Mentioning somebody (including the speaker) did or is doing
    something specific and objective
14 - Quoting quantities, statistics, and data
15 - Claiming a correlation or causation
16 - Assertion of existing laws or rules of operation
17 - Pledging a specific future plan or making specific predictions
    about future
18
19 ### Opinion
20
21 An opinion is a judgment based on facts, an attempt to draw a
    reasonable conclusion from factual evidence.
22 While the underlying facts can be verified, the derived opinion
    remains subjective and is not universally verifiable.
23
24 ## Dimension 2: Misleadingness
25
26 Assess how misleading the tweet is, i.e., its potential to express
    misinformation, regardless of intent.
27
28 ### Definition
29
30 Misinformation refers to false or misleading information that is
    contradicted by empirical evidence or inconsistent with
    common or expert consensus, without imposing assumptions on
    the intent or the format of the content
31
32 Indicators of misinformation or misleadingness:
33 - It contains a factual error
34 - It contains a digitally altered photo or video
35 - It contains outdated information that may be misleading
36 - It is a misrepresentation or missing important context
37 - It presents an unverified claim as a fact
38 - It is a joke or satire that might be misinterpreted as a fact
39
40 ## Output Format
41
42 For each input tweet, return a single-line JSON string with the
    following structure:

```

```

43 |
44 | {
45 |   "type_scores": {
46 |     "claim": float (0 to 1),    # Degree to which the tweet presents
         a factual/verifiable claim
47 |     "opinion": float (0 to 1)  # Degree to which the tweet
         expresses a subjective opinion
48 |   },
49 |   "misleadingness": float (0 to 1) # How misleading the tweet is,
         where 0 = not misleading, 1 = extremely misleading
50 | }
51 |
52 | ## Guidelines:
53 |
54 | - "claim" and "opinion" scores are independent and do not need to
         sum to 1.
55 | - Only return the JSON string. Do not include any explanation or
         additional text.

```

S9.3 GPT Validation with PolitiFact

In addition to the manual validation conducted with student assistants, we further evaluate the performance of our GPT-based approach using expert-verified data from PolitiFact. Specifically, we collect all statements fact-checked by PolitiFact between July 15, 2024 and June 4, 2025, aligning with the time window of our dataset of requested posts. We obtain 1320 statements in total and retain those rated as (mostly) false (= 1016) or (mostly) true (= 49), resulting in 1065 statements for evaluation. Applying the same GPT-based annotation pipeline, we compare GPT assessments with PolitiFact's expert verdicts. GPT achieves a high performance (*Accuracy* = 0.801; *Weighted F1* = 0.857), consistent with the results of our manual validation. Notably, despite the substantial class imbalance toward false statements, GPT attains comparable recall for true (*Recall* = 0.816) and false statements (*Recall* = 0.800), demonstrating its effectiveness in identifying both accurate and misleading content.

S9.4 Likelihood of Receiving Community Notes for Political Posts

Fig. S9.2 presents the estimation results for political post regarding the likelihood of receiving community notes. The results remain robust and consistent with our main findings.

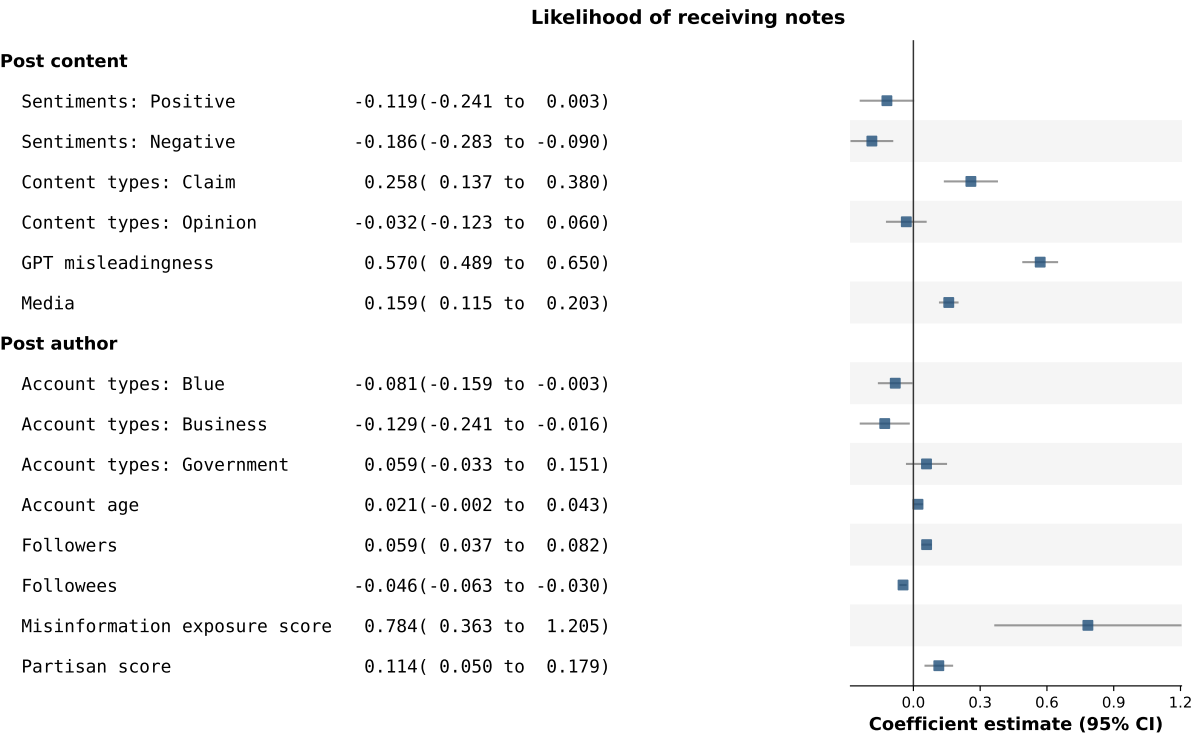


Figure S9.2: The estimation results for the logistic regression model predicting the likelihood that a requested political post receives a community note. Shown are coefficient estimates with 95% CIs. The coefficients for misinformation exposure score and partisan score are estimated based on the subset of posts containing the corresponding information. The number of words in each post is controlled during estimation but omitted in visualization for better readability. Continuous independent variables—word count, account age, number of followers, and number of followees—are z-standardized before estimation to facilitate interpretation.

S9.5 Robustness Check for Top Writers' Post Selection

In the main paper, we find that top writers are more likely to select political posts and posts with higher potential misleadingness after the request threshold, compared to before the request threshold. To ensure the robustness of our findings and reduce potential unobserved confounding related to potential heterogeneity among top writers, we restrict the analysis to posts selected by the same set of top writers before and after the request threshold. The estimation results are presented in Fig. S9.3. The estimates for political posts (*coef.* = 0.162, *odds ratio* = 1.176, $p < 0.001$) and misleadingness (*coef.* = 0.413, *odds ratio* = 1.511, $p < 0.001$) remain consistent with our findings, further supporting that requests shift contributors' attention toward political content and posts with higher misleadingness.

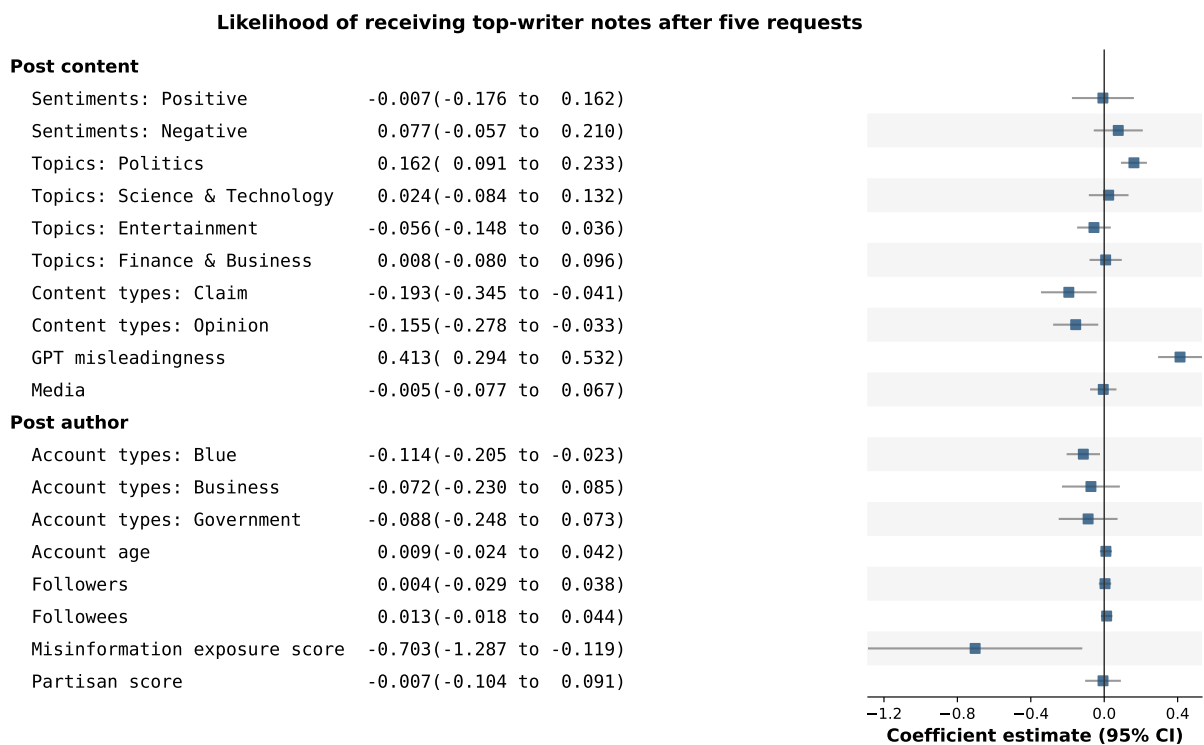


Figure S9.3: The estimation results for the logistic regression model predicting the likelihood that a requested post receives a top-writer note after the request threshold, *i.e.*, request-fostered notes. Shown are coefficient estimates with 95% CIs. The coefficients for misinformation exposure score and partisan score are estimated based on the subset of posts containing the corresponding information. The number of words in each post is controlled during estimation but omitted in visualization for better readability. Continuous independent variables—word count, account age, number of followers, and number of followees—are z-standardized before estimation to facilitate interpretation.

S9.6 Robustness Check for Likelihood of Receiving Community Notes

To further examine how Community Notes contributors differ from requestors in post selection and mitigate confounding from request-driven activity, we exclude posts with request-fostered notes identified in Section 9.4.2 and repeat our analysis on the likelihood of receiving community notes for the requested posts. The estimation results are reported in Fig. S9.4, and they remain robust and consistent with our main findings.

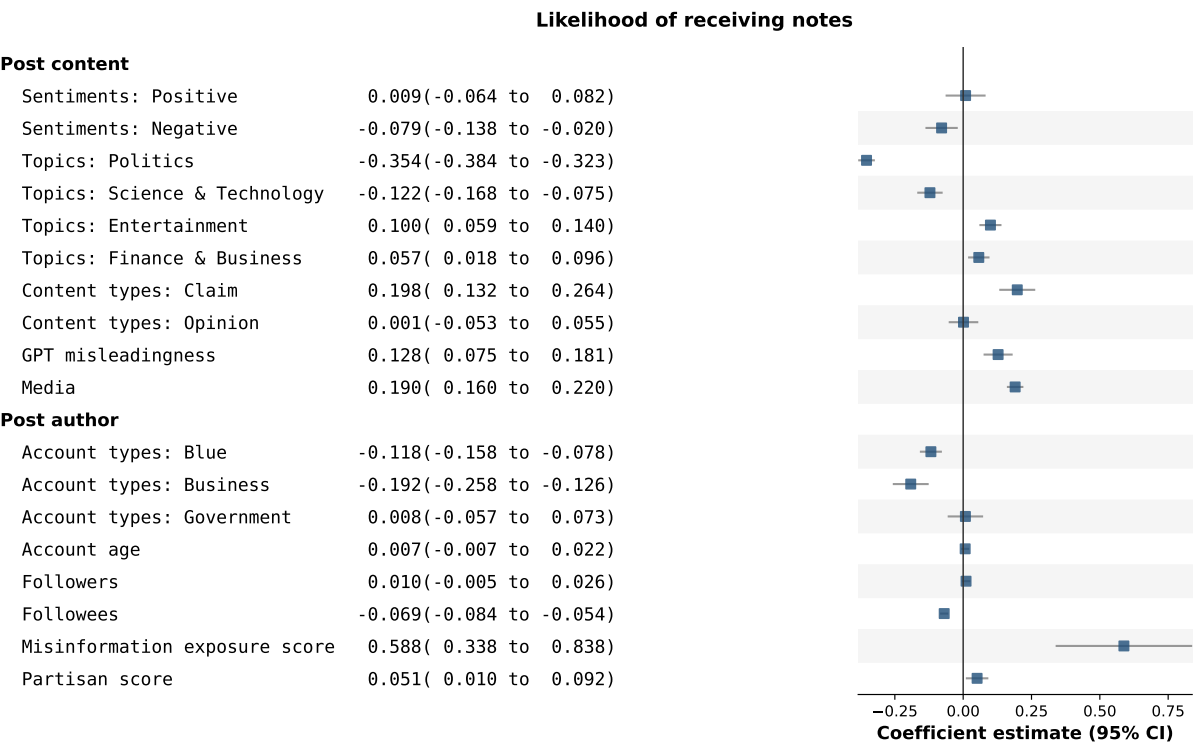


Figure S9.4: The estimation results for the logistic regression model predicting the likelihood that a requested post receives a community note. Shown are coefficient estimates with 95% Confidence Intervals (CIs). Posts with request-fostered notes are omitted during estimation. The number of words in each post is controlled during estimation but omitted in visualization for better readability. Continuous independent variables—word count, account age, number of followers, and number of followees—are z-standardized before estimation to facilitate interpretation.

Bibliography

- [1] U. K. Ecker et al., 'The psychological drivers of misinformation belief and its resistance to correction,' *Nature Reviews Psychology*, vol. 1, no. 1, pp. 13–29, 2022.
- [2] F. Pierri et al., 'Online misinformation is linked to early COVID-19 vaccination hesitancy and refusal,' *Scientific Reports*, vol. 12, no. 1, p. 5966, 2022.
- [3] A. M. Enders, J. Uscinski, C. Kloststad, and J. Stoler, 'On the relationship between conspiracy theory beliefs, misinformation, and vaccine hesitancy,' *PLoS ONE*, vol. 17, no. 10, e0276082, 2022.
- [4] J. Green, W. Hobbs, S. McCabe, and D. Lazer, 'Online engagement with 2020 election misinformation and turnout in the 2021 Georgia runoff election,' *Proceedings of the National Academy of Sciences*, vol. 119, no. 34, e2115900119, 2022.
- [5] D. M. J. Lazer et al., 'The science of fake news,' *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.
- [6] K. T. Gradoń, J. A. Hołyst, W. R. Moy, J. Sienkiewicz, and K. Suchecki, 'Countering misinformation: A multidisciplinary approach,' *Big Data & Society*, vol. 8, no. 1, p. 20 539 517 211 013 848, 2021.
- [7] L. Wu, F. Morstatter, K. M. Carley, and H. Liu, 'Misinformation in social media: Definition, manipulation, and detection,' *SIGKDD Explorations Newsletter*, vol. 21, no. 2, pp. 80–90, 2019.
- [8] I. Freiling, N. M. Krause, D. A. Scheufele, and D. Brossard, 'Believing and sharing misinformation, fact-checks, and accurate information on social media: The role of anxiety during COVID-19,' *New Media & Society*, vol. 25, no. 1, pp. 141–162, 2023.
- [9] N. M. Krause, I. Freiling, and D. A. Scheufele, 'The “infodemic” infodemic: Toward a more nuanced understanding of truth-claims and the need for (not) combatting misinformation,' *The ANNALS of the American Academy of Political and Social Science*, vol. 700, no. 1, pp. 112–123, 2022.
- [10] Y. Chuai and J. Zhao, 'Anger can make fake news viral online,' *Frontiers in Physics*, vol. 10, p. 970 174, 2022.
- [11] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, 'Fake news detection on social media,' *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017.
- [12] M. Pérez-Escolar, D. Lilleker, and A. Tapia-Frade, 'A systematic literature review of the phenomenon of disinformation and misinformation,' *Media and Communication*, vol. 11, no. 2, pp. 76–87, 2023.
- [13] J. N. Druckman, 'Threats to science: Politicization, misinformation, and inequalities,' *The ANNALS of the American Academy of Political and Social Science*, vol. 700, no. 1, pp. 8–24, 2022.
- [14] S. Vosoughi, D. Roy, and S. Aral, 'The spread of true and false news online,' *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.
- [15] O. Papakyriakopoulos and E. 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*, 2022, pp. 2541–2551.
- [16] A. Mathur and G. P. Moschis, 'How do information sources shape voters' political views?: Comparing mainstream and social-media effects on Democrats, Republicans, and the undecided,' *Journal of Advertising Research*, vol. 62, no. 2, pp. 176–195, 2022.
- [17] M. Mosleh and D. G. Rand, 'Measuring exposure to misinformation from political elites on Twitter,' *Nature Communications*, vol. 13, no. 1, p. 7144, 2022.
- [18] P. Suci, *Bots and misinformation on social media could undermine the midterm elections*, <https://www.forbes.com/sites/petersuci/2022/11/02/bots-and-misinformation-on-social-media-could-undermine-the-midterm-elections/>, [Accessed: 2022-11-03].
- [19] R. Recuero, F. B. Soares, and A. Gruz, 'Hyperpartisanship, disinformation and political conversations on Twitter: The Brazilian presidential election of 2018,' in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, 2020, pp. 569–578.

- [20] M. M. Haque, M. Yousuf, A. S. Alam, P. Saha, S. I. Ahmed, and N. Hassan, 'Combating misinformation in bangladesh: Roles and responsibilities as perceived by journalists, fact-checkers, and users,' *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW2, pp. 1–32, 2020.
- [21] H. Allcott and M. Gentzkow, 'Social media and fake news in the 2016 election,' *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–236, 2017.
- [22] K. Rapoza, *Can 'fake news' impact the stock market?* <https://www.forbes.com/sites/kenrapoza/2017/02/26/can-fake-news-impact-the-stock-market/>, [Accessed: 2024-08-26].
- [23] J. Clarke, H. Chen, D. Du, and Y. J. Hu, 'Fake news investor attention, and market reaction,' *Information Systems Research*, vol. 32, no. 1, pp. 35–52, 2021.
- [24] S. C. Briand et al., 'Infodemics: A new challenge for public health,' *Cell*, vol. 184, no. 25, pp. 6010–6014, 2021.
- [25] J. Roozenbeek et al., 'Susceptibility to misinformation about COVID-19 around the world,' *Royal Society Open Science*, vol. 7, no. 10, p. 201199, 2020.
- [26] S. Sultana and S. R. Fussell, 'Dissemination, situated fact-checking, and social effects of misinformation among rural Bangladeshi villagers during the COVID-19 pandemic,' *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, 436:1–436:34, 2021.
- [27] S. Z. Akbar, A. Panda, D. Kukreti, A. Meena, and J. Pal, 'Misinformation as a window into prejudice: COVID-19 and the information environment in India,' *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW3, 249:1–249:28, 2021.
- [28] R. A. Varanasi, J. Pal, and A. Vashistha, 'Accost, accede, or amplify: Attitudes towards COVID-19 misinformation on WhatsApp in India,' in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–17.
- [29] M. Cinelli, G. De Francisci Morales, A. Galeazzi, W. Quattrociocchi, and M. Starnini, 'The echo chamber effect on social media,' *Proceedings of the National Academy of Sciences*, vol. 118, no. 9, e2023301118, 2021.
- [30] S. Altay, M. Berriche, and A. Acerbi, 'Misinformation on misinformation: Conceptual and methodological challenges,' *Social Media + Society*, vol. 9, no. 1, p. 20563051221150412, 2023.
- [31] C. Budak, B. Nyhan, D. M. Rothschild, E. Thorson, and D. J. Watts, 'Misunderstanding the harms of online misinformation,' *Nature*, vol. 630, no. 8015, pp. 45–53, 2024.
- [32] R. Hartman et al., 'Interventions to reduce partisan animosity,' *Nature Human Behaviour*, vol. 6, no. 9, pp. 1194–1205, 2022.
- [33] U. Ecker et al., 'Misinformation poses a bigger threat to democracy than you might think,' *Nature*, vol. 630, no. 8015, pp. 29–32, 2024.
- [34] N. Pröllochs, D. Bär, and S. Feuerriegel, 'Emotions explain differences in the diffusion of true vs. false social media rumors,' *Scientific Reports*, vol. 11, no. 1, p. 22721, 2021, 22721.
- [35] M. Falkenberg et al., 'Growing polarization around climate change on social media,' *Nature Climate Change*, vol. 12, no. 12, pp. 1114–1121, 2022.
- [36] V. Sehgal, A. Peshin, S. Afroz, and H. Farid, 'Mutual hyperlinking among misinformation peddlers,' *arXiv*, 2021.
- [37] A. Miani, T. Hills, and A. Bangerter, 'Interconnectedness and (in)coherence as a signature of conspiracy worldviews,' *Science Advances*, vol. 8, no. 43, eabq3668, 2022.
- [38] X. Wang, M. Zhang, W. Fan, and K. Zhao, 'Understanding the spread of COVID-19 misinformation on social media: The effects of topics and a political leader's nudge,' *Journal of the Association for Information Science and Technology*, vol. 73, no. 5, pp. 726–737, 2022.
- [39] K. M. Douglas and R. M. Sutton, 'Why conspiracy theories matter: A social psychological analysis,' *European Review of Social Psychology*, vol. 29, no. 1, pp. 256–298, 2018.
- [40] Y. Kou, X. Gui, Y. Chen, and K. Pine, 'Conspiracy talk on social media: Collective sensemaking during a public health crisis,' *Proceedings of the ACM on Human-Computer Interaction*, vol. 1, no. CSCW, pp. 1–21, 2017.
- [41] V. Suarez-Lledo and J. Alvarez-Galvez, 'Prevalence of health misinformation on social media: Systematic review,' *Journal of Medical Internet Research*, vol. 23, no. 1, e17187, 2021.

- [42] H. W. A. Hanley, D. Kumar, and Z. Durumeric, 'A golden age: Conspiracy theories' relationship with misinformation outlets, news media, and the wider Internet,' *Proceedings of the ACM on Human-Computer Interaction*, vol. 7, no. CSCW2, pp. 1–33, 2023.
- [43] B. He, M. Ahamad, and S. Kumar, 'Reinforcement learning-based counter-misinformation response generation: A case study of COVID-19 vaccine misinformation,' in *Proceedings of the ACM Web Conference 2023*, 2023, pp. 2698–2709.
- [44] C. Martel and D. G. Rand, 'Misinformation warning labels are widely effective: A review of warning effects and their moderating features,' *Current Opinion in Psychology*, vol. 54, p. 101710, 2023.
- [45] I. Augenstein et al., 'Community moderation and the new epistemology of fact checking on social media,' *ArXiv*, 2025.
- [46] J. Allen, A. A. Arechar, G. Pennycook, and D. G. Rand, 'Scaling up fact-checking using the wisdom of crowds,' *Science Advances*, vol. 7, no. 36, eabf4393, 2021.
- [47] W. Godel et al., 'Moderating with the mob: Evaluating the efficacy of real-time crowdsourced fact-checking,' *Journal of Online Trust and Safety*, vol. 1, no. 1, pp. 1–36, 2021.
- [48] D. Dailey and K. Starbird, 'Journalists as crowdsourcerers: Responding to crisis by reporting with a crowd,' *Computer Supported Cooperative Work*, vol. 23, no. 4, pp. 445–481, 2014.
- [49] E. Agapie, J. Teevan, and A. Monroy-Hernández, 'Crowdsourcing in the field: A case study using local crowds for event reporting,' in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 3, 2015, pp. 2–11.
- [50] J. Allen, C. Martel, and D. G. Rand, 'Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program,' in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–19.
- [51] M. M. Bhuiyan, A. X. Zhang, C. M. Sehat, and T. Mitra, 'Investigating differences in crowdsourced news credibility assessment: Raters, tasks, and expert criteria,' *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW2, pp. 1–26, 2020.
- [52] Z. Epstein, G. Pennycook, and D. Rand, 'Will the crowd game the algorithm? Using layperson judgments to combat misinformation on social media by downranking distrusted sources,' in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–11.
- [53] G. Pennycook and D. G. Rand, 'Fighting misinformation on social media using crowdsourced judgments of news source quality,' *Proceedings of the National Academy of Sciences*, vol. 116, no. 7, pp. 2521–2526, 2019.
- [54] M. Saeed, N. Traub, M. Nicolas, G. Demartini, and P. Papotti, 'Crowdsourced fact-checking at Twitter: How does the crowd compare with experts?' In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 1736–1746.
- [55] Y. Zhang et al., 'Profiling the dynamics of trust & distrust in social media: A survey study,' in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–24.
- [56] C. P. Drolsbach, K. Solovev, and N. Pröllochs, 'Community notes increase trust in fact-checking on social media,' *PNAS Nexus*, vol. 3, no. 7, pgae217, 2024.
- [57] C. Chen and K. Shu, 'Combating misinformation in the age of LLMs: Opportunities and challenges,' *AI Magazine*, vol. 45, no. 3, pp. 354–368, 2024.
- [58] M. Glockner, Y. Hou, and I. Gurevych, 'Missing counter-evidence renders NLP fact-checking unrealistic for misinformation,' in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 5916–5936.
- [59] G. Warren, I. Shklovski, and I. Augenstein, 'Show me the work: Fact-checkers' requirements for explainable automated fact-checking,' in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025, pp. 1–21.
- [60] E. C. Choi and E. Ferrara, 'FACT-GPT: Fact-checking augmentation via claim matching with LLMs,' in *Companion Proceedings of the ACM Web Conference 2024*, 2024, pp. 883–886.
- [61] M. R. DeVerna, H. Y. Yan, K.-C. Yang, and F. Menczer, 'Fact-checking information from large language models can decrease headline discernment,' *Proceedings of the National Academy of Sciences*, vol. 121, no. 50, e2322823121, 2024.

- [62] L. Costabile, G. M. Orlando, V. La Gatta, and V. Moscato, 'Assessing the potential of generative agents in crowdsourced fact-checking,' *ArXiv*, 2025.
- [63] S. Rathje, D.-M. Mirea, I. Sucholutsky, R. Marjeh, C. E. Robertson, and J. J. Van Bavel, 'Gpt is an effective tool for multilingual psychological text analysis,' *Proceedings of the National Academy of Sciences*, vol. 121, no. 34, e2308950121, 2024.
- [64] L. Majer and J. Šnajder, 'Claim check-worthiness detection: How well do llms grasp annotation guidelines?' In *Proceedings of the Seventh Fact Extraction and Verification Workshop*, 2024, pp. 245–263.
- [65] Z. Lu, P. Li, W. Wang, and M. Yin, 'Understanding the effects of ai-based credibility indicators when people are influenced by both peers and experts,' in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025, pp. 1–19.
- [66] X Community Notes, *Signing up*, <https://communitynotes.x.com/guide/en/contributing/signing-up>, [Accessed: 2025-05-20].
- [67] S. Wojcik et al., 'Birdwatch: Crowd wisdom and bridging algorithms can inform understanding and reduce the spread of misinformation,' *ArXiv*, 2022.
- [68] R. A. Kaufman, M. R. Haupt, and S. P. Dow, 'Who's in the crowd matters: Cognitive factors and beliefs predict misinformation assessment accuracy,' *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW2, pp. 1–18, 2022.
- [69] T. Draws et al., 'The effects of crowd worker biases in fact-checking tasks,' in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 2114–2124.
- [70] M. R. Allen et al., 'Characteristics of X (formerly Twitter) Community Notes addressing COVID-19 vaccine misinformation,' *JAMA*, vol. 331, no. 19, pp. 1670–1672, 2024.
- [71] T. Kriplean, C. Bonnar, A. Borning, B. Kinney, and B. Gill, 'Integrating on-demand fact-checking with public dialogue,' in *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 2014, pp. 1188–1199.
- [72] Y. Chuai, J. Zhao, and G. Lenzini, 'From news source sharers to post viewers: How topic diversity and conspiracy theories shape engagement with misinformation during a health crisis,' *Proceedings of the ACM on Human-Computer Interaction*, vol. 9, no. 7, pp. 1–53, 2025.
- [73] Y. Chuai, J. Zhao, N. Pröllochs, and G. Lenzini, 'Is fact-checking politically neutral? Asymmetries in how U.S. fact-checking organizations pick up false statements mentioning political elites,' in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 19, 2025, pp. 403–429.
- [74] Y. Chuai, H. Tian, N. Pröllochs, and G. Lenzini, 'Did the roll-out of Community Notes reduce engagement with misinformation on X/Twitter?' *Proceedings of the ACM on Human-Computer Interaction*, vol. 8, no. CSCW2, pp. 1–52, 2024.
- [75] Y. Chuai, M. Horta Ribeiro, G. Lenzini, and N. Pröllochs, 'When "likers" go private: Engagement with reputationally risky content on X,' *ArXiv*, 2026.
- [76] Y. Chuai et al., 'Community-based fact-checking reduces the spread of misleading posts on X (formerly Twitter),' *Nature Communications*, 2026, forthcoming.
- [77] Y. Chuai, A. Sergeeva, G. Lenzini, and N. Pröllochs, 'Community fact-checks trigger moral outrage in replies to misleading posts on social media,' in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025, pp. 1–23.
- [78] Y. Chuai, G. Lenzini, and N. Pröllochs, 'Consensus stability of community notes on X,' in *Proceedings of the ACM on Web Conference 2026*, 2026, pp. 8885–8896.
- [79] Y. Chuai, S. Zhang, Z. Wang, X. Yi, M. Mosleh, and G. Lenzini, 'Request a note: How the request function shapes X's Community Notes system,' in *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 2026, pp. 1–22.
- [80] C. Chen, H. Wang, M. Shapiro, Y. Xiao, F. Wang, and K. Shu, 'Combating health misinformation in social media: Characterization, detection, intervention, and open issues,' *ArXiv*, 2022.
- [81] K. Solovev and N. Pröllochs, 'Moral emotions shape the virality of COVID-19 misinformation on social media,' in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 3706–3717.
- [82] Y. Chuai, Y. Chang, and J. Zhao, 'What really drives the spread of COVID-19 tweets: A revisit from perspective of content,' in *2022 IEEE 9th International Conference on Data Science and Advanced Analytics*, 2022, pp. 1–10.

- [83] L. L. Chen, S. R. Wilson, S. Lohmann, and D. V. Negraia, 'What are you anxious about? Examining subjects of anxiety during the COVID-19 pandemic,' in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 17, 2023, pp. 137–148.
- [84] D. Erokhin, A. Yosipof, and N. Komendantova, 'COVID-19 conspiracy theories discussion on Twitter,' *Social Media + Society*, vol. 8, no. 4, p. 20 563 051 221 126 051, 2022.
- [85] K. Sharma, Y. Zhang, and Y. Liu, 'COVID-19 vaccine misinformation campaigns and social media narratives,' in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, 2022, pp. 920–931.
- [86] X. Teng, Y.-R. Lin, W.-T. Chung, A. Li, and A. Kovashka, 'Characterizing user susceptibility to COVID-19 misinformation on Twitter,' in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, 2022, pp. 1005–1016.
- [87] R. Chandrasekaran, V. Mehta, T. Valkunde, and E. Moustakas, 'Topics, trends, and sentiments of tweets about the COVID-19 pandemic: Temporal infoveillance study,' *Journal of Medical Internet Research*, vol. 22, no. 10, e22624, 2020.
- [88] F. Cinus, M. Minici, C. Monti, and F. Bonchi, 'The effect of people recommenders on echo chambers and polarization,' in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, 2022, pp. 90–101.
- [89] K. Engel, S. Phadke, and T. Mitra, 'Learning from the ex-believers: Individuals' journeys in and out of conspiracy theories online,' *Proceedings of the ACM on Human-Computer Interaction*, vol. 7, no. CSCW2, pp. 1–37, 2023.
- [90] S. Lee, H. Seo, D. Lee, and A. Xiong, 'Associative inference can increase people's susceptibility to misinformation,' in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 17, 2023, pp. 530–541.
- [91] S. Prochaska et al., 'Mobilizing manufactured reality: How participatory disinformation shaped deep stories to catalyze action during the 2020 U.S. presidential election,' *Proceedings of the ACM on Human-Computer Interaction*, vol. 7, no. CSCW1, pp. 1–39, 2023.
- [92] J.-W. Van Prooijen, K. M. Douglas, and C. De Inocencio, 'Connecting the dots: Illusory pattern perception predicts belief in conspiracies and the supernatural,' *European Journal of Social Psychology*, vol. 48, no. 3, pp. 320–335, 2018.
- [93] M. Schmitz, G. Muric, and K. Burghardt, 'Detecting anti-vaccine users on Twitter,' in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 17, 2023, pp. 787–795.
- [94] S. Kim and J. Kim, 'The information ecosystem of conspiracy theory: Examining the QAnon narrative on Facebook,' *Proceedings of the ACM on Human-Computer Interaction*, vol. 7, no. CSCW1, pp. 1–24, 2023.
- [95] C. Ballard et al., 'Conspiracy brokers: Understanding the monetization of YouTube conspiracy theories,' in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 2707–2718.
- [96] C. Drolsbach and N. Pröllochs, 'Diffusion of community fact-checked misinformation on Twitter,' *Proceedings of the ACM on Human-Computer Interaction*, vol. 7, no. CSCW2, pp. 1–22, 2023.
- [97] C. G. Horner, D. Galletta, J. Crawford, and A. Shirsat, 'Emotions: The unexplored fuel of fake news on social media,' *Journal of Management Information Systems*, vol. 38, no. 4, pp. 1039–1066, 2021.
- [98] C. E. Robertson, N. Pröllochs, K. Schwarzenegger, P. Pärnamets, J. J. Van Bavel, and S. Feuerriegel, 'Negativity drives online news consumption,' *Nature Human Behaviour*, vol. 7, no. 5, pp. 812–822, 2023.
- [99] R. K. Garrett and R. M. Bond, 'Conservatives' susceptibility to political misperceptions,' *Science Advances*, vol. 7, no. 23, eabf1234, 2021.
- [100] A. Rao, F. Morstatter, and K. Lerman, 'Partisan asymmetries in exposure to misinformation,' *Scientific Reports*, vol. 12, no. 1, p. 15 671, 2022.
- [101] R. E. Robertson, J. Green, D. J. Ruck, K. Ognyanova, C. Wilson, and D. Lazer, 'Users choose to engage with more partisan news than they are exposed to on Google Search,' *Nature*, pp. 1–7, 2023.
- [102] M. Osmundsen, A. Bor, P. B. Vahlstrup, A. Bechmann, and M. B. Petersen, 'Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter,' *American Political Science Review*, vol. 115, no. 3, pp. 999–1015, 2021.
- [103] L. Jenke, 'Affective polarization and misinformation belief,' *Political Behavior*, pp. 1–60, 2023.

- [104] D. Nikolov, A. Flammini, and F. Menczer, 'Right and left, partisanship predicts (asymmetric) vulnerability to misinformation,' *Harvard Kennedy School Misinformation Review*, 2021.
- [105] S. Rathje, J. J. Van Bavel, and S. van der Linden, 'Out-group animosity drives engagement on social media,' *Proceedings of the National Academy of Sciences*, vol. 118, no. 26, e2024292118, 2021.
- [106] S. Lewandowsky, 'Climate change disinformation and how to combat it,' *Annual Review of Public Health*, vol. 42, pp. 1–21, 2021.
- [107] H. Fischer, M. Huff, and N. Said, 'Polarized climate change beliefs: No evidence for science literacy driving motivated reasoning in a U.S. national study,' *American Psychologist*, vol. 77, pp. 822–835, 2022.
- [108] J. Farrell, 'The growth of climate change misinformation in US philanthropy: Evidence from natural language processing,' *Environmental Research Letters*, vol. 14, no. 3, p. 034 013, 2019.
- [109] S. Loomba, A. de Figueiredo, S. J. Piatek, K. de Graaf, and H. J. Larson, 'Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA,' *Nature Human Behaviour*, vol. 5, no. 3, pp. 337–348, 2021.
- [110] N. Pröllochs, D. Bär, and S. Feuerriegel, 'Emotions in online rumor diffusion,' *EPJ Data Science*, vol. 10, no. 1, p. 51, 2021.
- [111] M. Charquero-Ballester, J. G. Walter, I. A. Nissen, and A. Bechmann, 'Different types of COVID-19 misinformation have different emotional valence on Twitter,' *Big Data & Society*, vol. 8, no. 2, p. 20 539 517 211 041 279, 2021.
- [112] Y. Wang, C. Ling, and G. Stringhini, 'Understanding the use of images to spread COVID-19 misinformation on Twitter,' *Proceedings of the ACM on Human-Computer Interaction*, vol. 7, no. CSCW1, pp. 1–32, 2023.
- [113] C. S. B. Ngai, R. G. Singh, and L. Yao, 'Impact of COVID-19 vaccine misinformation on social media virality: Content analysis of message themes and writing strategies,' *Journal of Medical Internet Research*, vol. 24, no. 7, e37806, 2022.
- [114] J. T. Huang, M. Krupenkin, D. Rothschild, and J. Lee Cunningham, 'The cost of anti-Asian racism during the COVID-19 pandemic,' *Nature Human Behaviour*, vol. 7, no. 5, pp. 682–695, 2023.
- [115] N. Velásquez et al., 'Online hate network spreads malicious COVID-19 content outside the control of individual social media platforms,' *Scientific Reports*, vol. 11, no. 1, p. 11 549, 2021.
- [116] M. Luca, E. Pronkina, and M. Rossi, 'The evolution of discrimination in online markets: How the rise in anti-Asian bias affected Airbnb during the pandemic,' *Marketing Science*, 2024.
- [117] C. Tang, S. Li, Y. Ding, R. D. Gopal, and G. Zhang, 'Racial discrimination and anti-discrimination: The COVID-19 pandemic's impact on Chinese restaurants in North America,' *Information Systems Research*, vol. 35, no. 3, pp. 1274–1295, 2024.
- [118] Y. Wang, H. Zhang, W. Gao, and C. Yang, 'Spillover effects from news to travel and leisure stocks during the COVID-19 pandemic: Evidence from the time and frequency domains,' *Tourism Economics*, vol. 29, no. 2, pp. 460–487, 2023.
- [119] S. E. Kreps and D. L. Kriner, 'Model uncertainty, political contestation, and public trust in science: Evidence from the COVID-19 pandemic,' *Science Advances*, vol. 6, no. 43, eabd4563, 2020.
- [120] G. Davies, E. Wu, and R. Frank, 'A witch's brew of grievances: The potential effects of COVID-19 on radicalization to violent extremism,' *Studies in Conflict & Terrorism*, vol. 46, no. 11, pp. 2327–2350, 2023.
- [121] P. S. Hart, S. Chinn, and S. Soroka, 'Politicization and polarization in COVID-19 news coverage,' *Science Communication*, vol. 42, no. 5, pp. 679–697, 2020.
- [122] E. Dehghan and A. Nagappa, 'Politicization and radicalization of discourses in the alt-tech ecosystem: A case study on Gab Social,' *Social Media + Society*, vol. 8, no. 3, p. 20 563 051 221 113 075, 2022.
- [123] J. D. Moffitt, C. King, and K. M. Carley, 'Hunting conspiracy theories during the COVID-19 pandemic,' *Social Media + Society*, vol. 7, no. 3, p. 20 563 051 211 043 212, 2021.
- [124] Y. Li, B. Jiang, K. Shu, and H. Liu, 'MM-COVID: A multilingual and multimodal data repository for combating COVID-19 disinformation,' *ArXiv*, 2020.

- [125] B. R. Taira, V. Kreger, A. Orue, and L. C. Diamond, 'A pragmatic assessment of Google Translate for emergency department instructions,' *Journal of General Internal Medicine*, vol. 36, no. 11, pp. 3361–3365, 2021.
- [126] P. Sebo and S. de Lucia, 'Performance of machine translators in translating French medical research abstracts to English: A comparative study of DeepL, Google Translate, and CUBBITT,' *PLoS ONE*, vol. 19, no. 2, e0297183, 2024.
- [127] C. Zhou, K. Li, and Y. Lu, 'Linguistic characteristics and the dissemination of misinformation in social media: The moderating effect of information richness,' *Information Processing & Management*, vol. 58, no. 6, p. 102679, 2021.
- [128] R. Plutchik, *The emotions*. University Press of America, 1991.
- [129] A. Papasavva et al., 'The gospel according to Q: Understanding the QAnon conspiracy from the perspective of canonical information,' in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, 2022, pp. 735–746.
- [130] D. Bär, N. Pröllochs, and S. Feuerriegel, 'Finding Qs: Profiling QAnon supporters on Parler,' in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 17, 2023, pp. 34–46.
- [131] R. C. Moore, R. Dahlke, P. L. Forberg, and J. T. Hancock, 'The private life of QAnon: A mixed methods investigation of Americans' exposure to QAnon content on the web,' *Proceedings of the ACM on Human-Computer Interaction*, vol. 8, no. CSCW2, pp. 1–34, 2024.
- [132] M. Hoseini, P. Melo, F. Benevenuto, A. Feldmann, and S. Zannettou, 'On the globalization of the QAnon conspiracy theory through Telegram,' in *Proceedings of the 15th ACM Web Science Conference 2023*, 2023, pp. 75–85.
- [133] K. Sharma, E. Ferrara, and Y. Liu, 'Characterizing online engagement with disinformation and conspiracies in the 2020 US presidential election,' in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, 2022, pp. 908–919.
- [134] M. Yan, Y.-R. Lin, and W.-T. Chung, 'Are mutated misinformation more contagious? A case study of COVID-19 misinformation on Twitter,' in *Proceedings of the 14th ACM Web Science Conference 2022*, 2022, pp. 336–347.
- [135] K. M. Douglas et al., 'Understanding conspiracy theories,' *Political Psychology*, vol. 40, pp. 3–35, 2019.
- [136] C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer, 'The spread of low-credibility content by social bots,' *Nature Communications*, vol. 9, no. 1, pp. 1–9, 2018.
- [137] S. Zannettou, '"I Won the Election!": An empirical analysis of soft moderation interventions on Twitter,' in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 15, 2021, pp. 865–876.
- [138] S. Pareek, N. van Berkel, E. Velloso, and J. Goncalves, 'Effect of explanation conceptualisations on trust in AI-assisted credibility assessment,' *Proceedings of the ACM on Human-Computer Interaction*, vol. 8, no. CSCW2, pp. 1–31, 2024.
- [139] L. Zavolokina, K. Sprenkamp, Z. Katashinskaya, D. G. Jones, and G. Schwabe, 'Think fast, think slow, think critical: Designing an automated propaganda detection tool,' in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–24.
- [140] F. Sharevski, R. Alsaadi, P. Jachim, and E. Pieroni, 'Misinformation warnings: Twitter's soft moderation effects on COVID-19 vaccine belief echoes,' *Computers & Security*, vol. 114, p. 102577, 2022.
- [141] D. Bär, N. Pröllochs, and S. Feuerriegel, 'New threats to society from free-speech social media platforms,' *Communications of the ACM*, vol. 66, no. 10, pp. 37–40, 2023.
- [142] M. Horta Ribeiro, J. Cheng, and R. West, 'Automated content moderation increases adherence to community guidelines,' in *Proceedings of the ACM Web Conference 2023*, 2023, pp. 2666–2676.
- [143] A. Flores et al., 'Politicians polarize and experts depolarize public support for COVID-19 management policies across countries,' *Proceedings of the National Academy of Sciences*, vol. 119, no. 3, e2117543119, 2022.
- [144] S. Chinn, P. S. Hart, and S. Soroka, 'Politicization and polarization in climate change news content, 1985-2017,' *Science Communication*, vol. 42, no. 1, pp. 112–129, 2020.
- [145] J. Shin and K. Thorson, 'Partisan selective sharing: The biased diffusion of fact-checking messages on social media,' *Journal of Communication*, vol. 67, no. 2, pp. 233–255, 2017.

- [146] B. Simon, K. M. Reininger, C. D. Schaefer, S. Zitzmann, and S. Kryś, 'Politicization as an antecedent of polarization: Evidence from two different political and national contexts,' *British Journal of Social Psychology*, vol. 58, no. 4, pp. 769–785, 2019.
- [147] C. Raymond, I. Waller, and A. Anderson, 'Measuring alignment of online grassroots political communities with political campaigns,' in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, 2022, pp. 806–816.
- [148] M. Saveski, D. Beeferman, D. McClure, and D. Roy, 'Engaging politically diverse audiences on social media,' in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, 2022, pp. 873–884.
- [149] R. Heiberger, S. Majó-Vázquez, L. Castro Herrero, R. K. Nielsen, and F. Esser, 'Do not blame the media! The role of politicians and parties in fragmenting online political debate,' *The International Journal of Press/Politics*, vol. 27, no. 4, pp. 910–941, 2022.
- [150] M. Pilarski, K. O. Solovev, and N. Pröllochs, 'Community notes vs. snoping: How the crowd selects fact-checking targets on social media,' in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 18, 2024, pp. 1262–1275.
- [151] C. M. Sehat, R. Li, P. Nie, T. Prabhakar, and A. X. Zhang, 'Misinformation as a harm: Structured approaches for fact-checking prioritization,' *Proceedings of the ACM on Human-Computer Interaction*, vol. 8, no. CSCW1, pp. 1–36, 2024.
- [152] S. Lee, A. Xiong, H. Seo, and D. Lee, '"Fact-checking" fact checkers: A data-driven approach,' *Harvard Kennedy School Misinformation Review*, 2023.
- [153] S. Nieminen and L. Rapeli, 'Fighting misperceptions and doubting journalists' objectivity: A review of fact-checking literature,' *Political Studies Review*, vol. 17, no. 3, pp. 296–309, 2019.
- [154] S. Singh, S. Masud, and T. Chakraborty, 'Independent fact-checking organizations exhibit a departure from political neutrality,' *ArXiv*, 2024.
- [155] S. Feuerriegel et al., 'Research can help to tackle AI-generated disinformation,' *Nature Human Behaviour*, vol. 7, no. 11, pp. 1818–1821, 2023.
- [156] N. Micallef, V. Armacost, N. Memon, and S. Patil, 'True or false: Studying the work practices of professional fact-checkers,' *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW1, pp. 1–44, 2022.
- [157] S. González-Bailón, V. d'Andrea, D. Freelon, and M. De Domenico, 'The advantage of the right in social media news sharing,' *PNAS Nexus*, vol. 1, no. 3, pgac137, 2022.
- [158] M. Grossmann and D. A. Hopkins, *Asymmetric politics: Ideological Republicans and group interest Democrats*. Oxford University Press, 2016.
- [159] E. Ostermeier, *Selection bias? PolitiFact rates Republican statements as false at three times the rate of Democrats*, <https://smartpolitics.lib.umn.edu/2011/02/10/selection-bias-politifact-rate/>, [Accessed: 2024-08-14].
- [160] M. Dastani and A. Pankov, 'Other-condemning moral emotions: Anger, contempt and disgust,' *ACM Transactions on Internet Technology*, vol. 17, no. 1, pp. 1–24, 2017.
- [161] D. M. Markowitz, T. R. Levine, K. B. Serota, and A. D. Moore, 'Cross-checking journalistic fact-checkers: The role of sampling and scaling in interpreting false and misleading statements,' *PLoS ONE*, vol. 18, no. 7, e0289004, 2023.
- [162] S. Feuerriegel et al., 'Using natural language processing to analyse text data in behavioural science,' *Nature Reviews Psychology*, vol. forthcoming, 2025.
- [163] S. M. Mohammad and P. D. Turney, 'Crowdsourcing a word–emotion association lexicon,' *Computational Intelligence*, vol. 29, no. 3, pp. 436–465, 2013.
- [164] W. Chen, D. Pacheco, K.-C. Yang, and F. Menczer, 'Neutral bots probe political bias on social media,' *Nature Communications*, vol. 12, no. 1, p. 5580, 2021.
- [165] P. Mena, 'Principles and boundaries of fact-checking: Journalists' perceptions,' *Journalism Practice*, vol. 13, no. 6, pp. 657–672, 2019.
- [166] E. Saltz, S. Barari, C. Leibowicz, and C. Wardle, 'Misinformation interventions are common, divisive, and poorly understood,' *Harvard Kennedy School Misinformation Review*, 2021.

- [167] N. Pröllochs, 'Community-based fact-checking on Twitter's Birdwatch platform,' in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, 2022, pp. 794–805.
- [168] Twitter, *Helpful Birdwatch notes are now visible to everyone on Twitter in the US*, https://blog.twitter.com/en_us/topics/product/2022/helpful-birdwatch-notes-now-visible-everyone-twitter-us, [Accessed: 2023-04-26].
- [169] G. Pennycook and D. G. Rand, 'Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation,' *Nature Communications*, vol. 13, no. 1, p. 2333, 2022.
- [170] G. Pennycook, J. McPhetres, Y. Zhang, J. G. Lu, and D. G. Rand, 'Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention,' *Psychological Science*, vol. 31, no. 7, pp. 770–780, 2020.
- [171] G. Pennycook, Z. Epstein, M. Mosleh, A. A. Arechar, D. Eckles, and D. G. Rand, 'Shifting attention to accuracy can reduce misinformation online,' *Nature*, vol. 592, no. 7855, pp. 590–595, 2021.
- [172] J. Stein, V. Frey, and A. van de Rijt, 'Realtime user ratings as a strategy for combatting misinformation: An experimental study,' *Scientific Reports*, vol. 13, no. 1, p. 1626, 2023.
- [173] T. Yasseri and F. Menczer, 'Can crowdsourcing rescue the social marketplace of ideas?' *Communications of the ACM*, vol. 66, no. 9, pp. 42–45, 2023.
- [174] TikTok, *Taking action against COVID-19 vaccine misinformation*, <https://newsroom.tiktok.com/en-gb/taking-action-against-covid-19-vaccine-misinformation>, [Accessed: 2023-04-29].
- [175] Google, *Fighting disinformation across our products*, <https://blog.google/around-the-globe/google-europe/fighting-disinformation-across-our-products/>, [Accessed: 2025-05-20].
- [176] Meta, *Meta's third-party fact-checking program*, <https://www.facebook.com/formedia/mjp/programs/third-party-fact-checking>, [Accessed: 2023-05-02].
- [177] P. Juneja and T. Mitra, 'Human and technological infrastructures of fact-checking,' *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW2, pp. 1–36, 2022.
- [178] Twitter, *Community Notes*, <https://communitynotes.twitter.com/guide/en.html>, [Accessed: 2023-04-27].
- [179] C. P. Drolsbach and N. Pröllochs, 'Believability and harmfulness shape the virality of misleading social media posts,' in *Proceedings of the ACM Web Conference 2023*, 2023, pp. 4172–4177.
- [180] J. Pfeffer, D. Matter, and A. Sargsyan, 'The half-life of a tweet,' in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 17, 2023, pp. 1163–1167.
- [181] F. Jahanbakhsh, Y. Katsis, D. Wang, L. Popa, and M. Muller, 'Exploring the use of personalized AI for identifying misinformation on social media,' in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–27.
- [182] Y. Ma, B. He, N. Subrahmanian, and S. Kumar, 'Characterizing and predicting social correction on Twitter,' in *Proceedings of the 15th ACM Web Science Conference 2023*, 2023, pp. 86–95.
- [183] W. J. Brady, J. A. Wills, J. T. Jost, J. A. Tucker, and J. J. Van Bavel, 'Emotion shapes the diffusion of moralized content in social networks,' *Proceedings of the National Academy of Sciences*, vol. 114, no. 28, pp. 7313–7318, 2017.
- [184] J. D. Angrist and J.-S. Pischke, *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press, 2009.
- [185] M. Horta Ribeiro, H. Hosseinmardi, R. West, and D. J. Watts, 'Deplatforming did not decrease Parler users' activity on fringe social media,' *PNAS Nexus*, vol. 2, no. 3, pgad035, 2023.
- [186] Y. Liu, Y. Lu, K. Nayak, F. Zhang, L. Zhang, and Y. Zhao, 'Empirical analysis of EIP-1559: Transaction fees, waiting times, and consensus security,' in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2022, pp. 2099–2113.
- [187] G. L. S. Freire, T. Panoutsos, L. Perez, F. Benevenuto, and F. Figueiredo, 'Understanding effects of moderation and migration on online video sharing platforms,' in *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, 2022, pp. 220–224.
- [188] K. Conger, *How Elon Musk is changing the Twitter experience*, <https://www.nytimes.com/2023/04/07/technology/elon-musk-twitter-changes.html>, [Accessed: 2023-06-07].

- [189] L. Courchesne, J. Ilhardt, and J. N. Shapiro, 'Review of social science research on the impact of countermeasures against influence operations,' *Harvard Kennedy School Misinformation Review*, 2021.
- [190] K. Clayton et al., 'Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media,' *Political Behavior*, vol. 42, pp. 1073–1095, 2020.
- [191] G. Orosz, B. Paskuj, L. Faragó, and P. Krekó, 'A prosocial fake news intervention with durable effects,' *Scientific Reports*, vol. 13, no. 1, p. 3958, 2023.
- [192] I. Munyaka, E. Hargittai, and E. Redmiles, 'The misinformation paradox: Older adults are cynical about news media, but engage with it anyway,' *Journal of Online Trust and Safety*, vol. 1, no. 4, 2022.
- [193] G. Ceylan, I. A. Anderson, and W. Wood, 'Sharing of misinformation is habitual, not just lazy or biased,' *Proceedings of the National Academy of Sciences*, vol. 120, no. 4, e2216614120, 2023.
- [194] Z. Epstein, N. Sirlin, A. Arechar, G. Pennycook, and D. Rand, 'The social media context interferes with truth discernment,' *Science Advances*, vol. 9, no. 9, eabo6169, 2023.
- [195] T. Bucher, A. Helmond, et al., 'The affordances of social media platforms,' *The SAGE Handbook of Social Media*, vol. 1, no. 1, pp. 233–253, 2018.
- [196] A. Ronzhyn, A. S. Cardenal, and A. Batlle Rubio, 'Defining affordances in social media research: A literature review,' *New Media & Society*, vol. 25, no. 11, pp. 3165–3188, 2023.
- [197] M. A. DeVito, J. Birnholtz, and J. T. Hancock, 'Platforms, people, and perception: Using affordances to understand self-presentation on social media,' in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 2017, pp. 740–754.
- [198] B. Marder, E. Slade, D. Houghton, and C. Archer-Brown, "'I like them, but won't 'like' them": An examination of impression management associated with visible political party affiliation on Facebook,' *Computers in Human Behavior*, vol. 61, pp. 280–287, 2016.
- [199] S. Das and A. Kramer, 'Self-censorship on Facebook,' in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 7, 2013, pp. 120–127.
- [200] O. Tenenboim, 'Comments, shares, or likes: What makes news posts engaging in different ways,' *Social Media + Society*, vol. 8, no. 4, p. 20563051221130282, 2022.
- [201] C. Gerlitz and A. Helmond, 'The like economy: Social buttons and the data-intensive web,' *New Media & Society*, vol. 15, no. 8, pp. 1348–1365, 2013.
- [202] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, 'The rise of social bots,' *Communications of the ACM*, vol. 59, no. 7, pp. 96–104, 2016.
- [203] L. Zhu and K. Lerman, 'Attention inequality in social media,' *ArXiv*, 2016.
- [204] J. W. Treem and P. M. Leonardi, 'Social media use in organizations: Exploring the affordances of visibility, editability, persistence, and association,' *Annals of the International Communication Association*, vol. 36, no. 1, pp. 143–189, 2013.
- [205] R. Scharlach and B. Hallinan, 'The value affordances of social media engagement features,' *Journal of Computer-Mediated Communication*, vol. 28, no. 6, zmad040, 2023.
- [206] F. Menczer, 'The spread of misinformation in social media,' in *Proceedings of the 25th International Conference Companion on World Wide Web*, 2016, pp. 717–717.
- [207] E. Aimeur, S. Amri, and G. Brassard, 'Fake news, disinformation and misinformation in social media: A review,' *Social Network Analysis and Mining*, vol. 13, no. 1, p. 30, 2023.
- [208] A. Matamoros-Fernández and J. Farkas, 'Racism, hate speech, and social media: A systematic review and critique,' *Television & New Media*, vol. 22, no. 2, pp. 205–224, 2021.
- [209] B. Mathew, R. Dutt, P. Goyal, and A. Mukherjee, 'Spread of hate speech in online social media,' in *Proceedings of the 10th ACM Conference on Web Science*, 2019, pp. 173–182.
- [210] A. Maarouf, N. Pröllochs, and S. Feuerriegel, 'The virality of hate speech on social media,' *Proceedings of the ACM on Human-Computer Interaction*, vol. 8, no. CSCW1, pp. 1–22, 2024.
- [211] R. Y. M. Wong, C. M. Cheung, B. Xiao, and J. B. Thatcher, 'Standing up or standing by: Understanding bystanders' proactive reporting responses to social media harassment,' *Information Systems Research*, vol. 32, no. 2, pp. 561–581, 2021.

- [212] T. Gillespie, 'Do not recommend? Reduction as a form of content moderation,' *Social Media + Society*, vol. 8, no. 3, p. 2056305122117552, 2022.
- [213] S. Zhang et al., 'Commenotes: Synthesizing organic comments to support community-based fact-checking,' *ArXiv*, 2025.
- [214] K. Beadle, M. Warner, and M. Vasek, "'Edit: I'm sorry for being offensive, this is getting downvoted and I feel terrible': Implicit social norms as governance in identity-based communities," *Proceedings of the ACM on Human-Computer Interaction*, vol. 9, no. 2, pp. 1–32, 2025.
- [215] E. Chandrasekharan, S. Jhaver, A. Bruckman, and E. Gilbert, 'Quarantined! Examining the effects of a community-wide moderation intervention on Reddit,' *ACM Transactions on Computer-Human Interaction*, vol. 29, no. 4, pp. 1–26, 2022.
- [216] M. Horta Ribeiro et al., 'Do platform migrations compromise content moderation? Evidence from r/the_donald and r/incels,' *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–24, 2021.
- [217] S. Jhaver, D. S. Appling, E. Gilbert, and A. Bruckman, "'Did you suspect the post would be removed?' Understanding user reactions to content removals on Reddit," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–33, 2019.
- [218] S. Jhaver, C. Boylston, D. Yang, and A. Bruckman, 'Evaluating the effectiveness of deplatforming as a moderation strategy on Twitter,' *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–30, 2021.
- [219] T. Mitts, N. Pisharody, and J. Shapiro, 'Removal of anti-vaccine content impacts social media discourse,' in *Proceedings of the 14th ACM Web Science Conference 2022*, 2022, pp. 319–326.
- [220] S. Ali et al., 'Understanding the effect of deplatforming on social networks,' in *Proceedings of the 13th ACM Web Science Conference 2021*, 2021, pp. 187–195.
- [221] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert, 'You can't stay here: The efficacy of Reddit's 2015 ban examined through hate speech,' *Proceedings of the ACM on Human-Computer Interaction*, vol. 1, no. CSCW, pp. 1–22, 2017.
- [222] S. Jhaver, A. Bruckman, and E. Gilbert, 'Does transparency in moderation really matter? User behavior after content removal explanations on reddit,' *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–27, 2019.
- [223] M. Horta Ribeiro, R. West, R. Lewis, and S. Kairam, 'Post guidance for online communities,' *Proceedings of the ACM on Human-Computer Interaction*, vol. 9, no. 2, pp. 1–26, 2025.
- [224] J. Kim and H. Zo, 'Am I watching or being watched? Exploring the selective disclosure paradox in users' self-censorship to dataveillance awareness in video recommender systems,' *Telematics and Informatics*, vol. 98, p. 102253, 2025.
- [225] K. Saha, P. Gupta, G. Mark, E. Kiciman, and M. De Choudhury, 'Observer effect in social media use,' in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–20.
- [226] J. Birnholtz, A. Kraus, W. Zheng, D. A. Moskowicz, K. Macapagal, and D. Gergle, 'Sensitive sharing on social media: Exploring willingness to disclose PrEP usage among adolescent males who have sex with males,' *Social Media + Society*, vol. 6, no. 3, p. 2056305120955176, 2020.
- [227] D. Kekülluoglu, K. Vaniea, and W. Magdy, 'Understanding privacy switching behaviour on Twitter,' in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–14.
- [228] D. Kekülluoglu, K. Vaniea, M. K. Wolters, and W. Magdy, 'Twitter has a binary privacy setting, are users aware of how it works?' *Proceedings of the ACM on Human-Computer Interaction*, vol. 7, no. CSCW1, pp. 1–18, 2023.
- [229] K. Zhang and R. Kizilcec, 'Anonymity in social media: Effects of content controversiality and social endorsement on sharing behavior,' in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, 2014, pp. 643–646.
- [230] X. Pan, Y. Hou, and Q. Wang, 'Are we braver in cyberspace? social media anonymity enhances moral courage,' *Computers in Human Behavior*, vol. 148, p. 107880, 2023.
- [231] S. Huang and S. Lin, 'Do more likes lead to more clicks? evidence from a field experiment on social advertising,' *Journal of Marketing*, vol. 89, no. 5, pp. 88–110, 2025.

- [232] S.-Y. Lee, S. S. Hansen, and J. K. Lee, 'What makes us click "like" on Facebook? Examining psychological, technological, and motivational factors on virtual endorsement,' *Computer Communications*, vol. 73, pp. 332–341, 2016.
- [233] T. W. Post, *The end of an era: Likes on X are anonymous now*, <https://www.washingtonpost.com/technology/2024/06/12/twitter-likes-hidden-private-x/>, 2024.
- [234] M. Sleeper, R. Balebako, S. Das, A. L. McConahy, J. Wiese, and L. F. Cranor, 'The post that wasn't: Exploring self-censorship on Facebook,' in *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, 2013, pp. 793–802.
- [235] H. Elmimouni et al., 'Exploring algorithmic resistance: Responses to social media censorship in activism,' *Proceedings of the ACM on Human-Computer Interaction*, vol. 9, no. 2, pp. 1–24, 2025.
- [236] SPLC, *Extremists, groups, & ideologies*, https://www.splcenter.org/resources/extremist-files/?meta_extremist_type=extremist, [Accessed: 26-05-2025].
- [237] Wikidata, *Wikidata query service*, <https://query.wikidata.org/>, [Accessed: 26-05-2025].
- [238] J. B. Lewis, K. Poole, H. Rosenthal, A. Boche, A. Rudkin, and L. Sonnet, *Voteview: Congressional roll-call votes database*, <https://voteview.com/data/>, [Accessed: 26-05-2025].
- [239] AllSides, *X influencer bias chart & ratings*, <https://www.allsides.com/media-bias/x-bias-chart>, [Accessed: 26-05-2025].
- [240] FeedSpot, *Top 100 tech influencers in 2025*, https://influencers.feedspot.com/tech_instagram_influencers/, [Accessed: 26-05-2025].
- [241] A. Rambachan and J. Roth, 'An honest approach to parallel trends,' *Unpublished manuscript, Harvard University*, 2019.
- [242] P. R. Center, *Americans' social media use*, <https://www.pewresearch.org/internet/2024/01/31/americans-social-media-use/>, [Accessed: 26-05-2025].
- [243] C.-Y. Chin, H.-P. Lu, and C.-M. Wu, 'Facebook users' motivation for clicking the "like" button,' *Social Behavior and Personality: An International Journal*, vol. 43, no. 4, pp. 579–592, 2015.
- [244] A. Levordashka, S. Utz, and R. Ambros, 'What's in a like? motivations for pressing the like button,' in *Proceedings of the International AAAI Conference on Web and Social media*, vol. 10, 2016, pp. 623–626.
- [245] T. L. Webb and P. Sheeran, 'Does changing behavioral intentions engender behavior change? a meta-analysis of the experimental evidence.,' *Psychological Bulletin*, vol. 132, no. 2, p. 249, 2006.
- [246] R. E. Rhodes and L. Dickau, 'Experimental evidence for the intention–behavior relationship in the physical activity domain: A meta-analysis.,' *Health Psychology*, vol. 31, no. 6, p. 724, 2012.
- [247] P. Sheeran and T. L. Webb, 'The intention–behavior gap,' *Social and Personality Psychology Compass*, vol. 10, no. 9, pp. 503–518, 2016.
- [248] K. Gligoric, J. Czeszchowska, A. Anderson, and R. West, 'Anticipated versus actual effects of platform design change: A case study of Twitter's character limit,' *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW2, pp. 1–29, 2022.
- [249] Q. Zhang, H. Liang, T.-Q. Peng, and J. J. Zhu, 'The effect of affordance on deliberation when retweeting: From the perspective of expression effect,' *Computers in Human Behavior*, vol. 151, p. 108 010, 2024.
- [250] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer, 'Fake news on Twitter during the 2016 U.S. presidential election,' *Science*, vol. 363, no. 6425, pp. 374–378, 2019.
- [251] S. Baribi-Bartov, B. Swire-Thompson, and N. Grinberg, 'Supersharers of fake news on Twitter,' *Science*, vol. 384, no. 6699, pp. 979–982, 2024.
- [252] K. Sharma, Y. Zhang, E. Ferrara, and Y. Liu, 'Identifying coordinated accounts on social media through hidden influence and group behaviours,' in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 1441–1451.
- [253] S. Aral and D. Eckles, 'Protecting elections from social media manipulation,' *Science*, vol. 365, no. 6456, pp. 858–861, 2019.
- [254] E. Bakshy, S. Messing, and L. A. Adamic, 'Exposure to ideologically diverse news and opinion on Facebook,' *Science*, vol. 348, no. 6239, pp. 1130–1132, 2015.

- [255] A. M. Guess, B. Nyhan, and J. Reifler, 'Exposure to untrustworthy websites in the 2016 US election,' *Nature Human Behaviour*, vol. 4, no. 5, pp. 472–480, 2020.
- [256] R. C. Moore, R. Dahlke, and J. T. Hancock, 'Exposure to untrustworthy websites in the 2020 US election,' *Nature Human Behaviour*, vol. 7, no. 7, pp. 1096–1105, 2023.
- [257] D. A. Broniatowski et al., 'Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate,' *American Journal of Public Health*, vol. 108, no. 10, pp. 1378–1384, 2018.
- [258] Y. M. Rocha, G. A. de Moura, G. A. Desidério, C. H. de Oliveira, F. D. Lourenço, and L. D. de Figueiredo Nicolete, 'The impact of fake news on social media and its influence on health during the COVID-19 pandemic: A systematic review,' *Journal of Public Health*, vol. 31, pp. 1007–1016, 2021.
- [259] R. Gallotti, F. Valle, N. Castaldo, P. Sacco, and M. De Domenico, 'Assessing the risks of 'infodemics' in response to COVID-19 epidemics,' *Nature Human Behaviour*, vol. 4, no. 12, pp. 1285–1293, 2020.
- [260] K. Starbird, 'Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on Twitter,' in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, 2017, pp. 230–239.
- [261] O. Oh, M. Agrawal, and H. R. Rao, 'Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises,' *MIS Quarterly*, vol. 37, no. 2, pp. 407–426, 2013.
- [262] R. Calo, C. Coward, E. S. Spiro, K. Starbird, and J. D. West, 'How do you solve a problem like misinformation?' *Science Advances*, vol. 7, no. 50, eabn0481, 2021.
- [263] J. Donovan, 'Social-media companies must flatten the curve of misinformation,' *Nature*, 2020.
- [264] A. Kozyreva et al., 'Toolbox of individual-level interventions against online misinformation,' *Nature Human Behaviour*, vol. 8, no. 6, pp. 1044–1052, 2024.
- [265] A. Mosseri, *Addressing hoaxes and fake news*, <https://about.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/>, 2016.
- [266] Instagram, *Combating misinformation on Instagram*, <https://about.instagram.com/blog/announcements/combating-misinformation-on-instagram>, 2019.
- [267] S. Altay, M. Berriche, H. Heuer, J. Farkas, and S. Rathje, 'A survey of expert views on misinformation: Definitions, determinants, solutions, and future of the field,' *HKS Misinformation Review*, vol. 4, no. 4, 2023.
- [268] A. Kim and A. R. Dennis, 'Says who? The effects of presentation format and source rating on fake news in social media,' *MIS Quarterly*, vol. 43, no. 3, pp. 1025–1039, 2019.
- [269] K. C. Ng, J. Tang, and D. Lee, 'The effect of platform intervention policies on fake news dissemination and survival: An empirical examination,' *Journal of Management Information Systems*, vol. 38, no. 4, pp. 898–930, 2021.
- [270] E. Porter and T. 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*, vol. 118, no. 37, e2104235118, 2021.
- [271] P. L. Moravec, A. Kim, and A. R. Dennis, 'Appealing to sense and sensibility: System 1 and system 2 interventions for fake news on social media,' *Information Systems Research*, vol. 31, no. 3, pp. 987–1006, 2020.
- [272] G. Pennycook, A. Bear, E. T. Collins, and D. G. Rand, 'The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings,' *Management Science*, vol. 66, no. 11, pp. 4944–4957, 2020.
- [273] P. Mena, 'Cleaning up social media: The effect of warning labels on likelihood of sharing false news on Facebook,' *Policy & Internet*, vol. 12, no. 2, pp. 165–183, 2020.
- [274] W. Yaqub, O. Kakhidze, M. L. Brockman, N. Memon, and S. Patil, 'Effects of credibility indicators on social media news sharing intent,' in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–14.
- [275] Z. Guo, M. Schlichtkrull, and A. Vlachos, 'A survey on automated fact-checking,' *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 178–206, 2022.

- [276] C. T. Robertson, R. R. Mourão, and E. Thorson, 'Who uses fact-checking sites? The impact of demographics, political antecedents, and media use on fact-checking site awareness, attitudes, and behavior,' *The International Journal of Press/Politics*, vol. 25, no. 2, pp. 217–237, 2020.
- [277] J. Straub and M. Spradling, 'Americans' perspectives on online media warning labels,' *Behavioral Sciences*, vol. 12, no. 3, p. 59, 2022.
- [278] D. Flamini, *Most Republicans don't trust fact-checkers, and most Americans don't trust the media*, <https://www.poynter.org/ifcn/2019/most-republicans-dont-trust-fact-checkers-and-most-americans-dont-trust-the-media/>, 2019.
- [279] N. Micallef, B. He, S. Kumar, M. Ahamad, and N. Memon, 'The role of the crowd in countering misinformation: A case study of the COVID-19 infodemic,' in *2020 IEEE International Conference on Big Data*, 2020, pp. 748–757.
- [280] J. Allen, B. Howland, M. Mobius, D. Rothschild, and D. J. Watts, 'Evaluating the fake news problem at the scale of the information ecosystem,' *Science Advances*, vol. 6, no. 14, 2020, eaay3539.
- [281] C. Martel, J. Allen, G. Pennycook, and D. G. Rand, 'Crowds can effectively identify misinformation at scale,' *Perspectives on Psychological Science*, vol. 19, no. 2, pp. 477–488, 2024.
- [282] V. Frey and A. van de Rijt, 'Social influence undermines the wisdom of the crowd in sequential decision making,' *Management Science*, vol. 67, no. 7, pp. 4273–4286, 2021.
- [283] A. W. Woolley, C. F. Chabris, A. Pentland, N. Hashmi, and T. W. Malone, 'Evidence for a collective intelligence factor in the performance of human groups,' *Science*, vol. 330, no. 6004, pp. 686–688, 2010.
- [284] C. Okoli, M. Mehdi, M. Mesgari, F. Å. Nielsen, and A. Lanamäki, 'Wikipedia in the eyes of its beholders: A systematic review of scholarly research on Wikipedia readers and readership,' *Journal of the Association for Information Science and Technology*, vol. 65, no. 12, pp. 2381–2403, 2014.
- [285] P. Resnick, A. Alfayez, J. Im, and E. Gilbert, 'Informed crowds can effectively identify misinformation,' *ArXiv*, 2021.
- [286] A. Zhao and M. Naaman, 'Insights from a comparative study on the variety, velocity, veracity, and viability of crowdsourced and professional fact-checking services,' *Journal of Online Trust and Safety*, vol. 2, no. 1, 2023.
- [287] X, *Introducing Birdwatch, a community-based approach to misinformation*, https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation, 2021.
- [288] K. Solovev and N. Pröllochs, 'References to unbiased sources increase the helpfulness of community fact-checks,' *Scientific Reports*, vol. 15, no. 1, p. 25749, 2025.
- [289] M. Del Vicario et al., 'The spreading of misinformation online,' *Proceedings of the National Academy of Sciences*, vol. 113, no. 3, pp. 554–559, 2016.
- [290] A. Friggeri, L. Adamic, D. Eckles, and J. Cheng, 'Rumor cascades,' in *proceedings of the international AAAI conference on web and social media*, vol. 8, 2014, pp. 101–110.
- [291] A. Bilinski and L. A. Hatfield, 'Nothing to see here? Non-inferiority approaches to parallel trends and other model assumptions,' *ArXiv*, 2018.
- [292] J. Roth, P. H. C. Sant'Anna, A. Bilinski, and J. Poe, 'What's trending in difference-in-differences? A synthesis of the recent econometrics literature,' *Journal of Econometrics*, vol. 235, no. 2, pp. 2218–2244, Aug. 2023.
- [293] D. Lakens, A. M. Scheel, and P. M. Isager, 'Equivalence testing for psychological research: A tutorial,' *Advances in Methods and Practices in Psychological Science*, vol. 1, no. 2, pp. 259–269, 2018.
- [294] A. Rambachan and J. Roth, 'A more credible approach to parallel trends,' *Review of Economic Studies*, vol. 90, no. 5, pp. 2555–2591, 2023.
- [295] X, *Twitter's recommendation algorithm*, https://blog.x.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm, 2023.
- [296] X, *Faqs for Community Notes program*, <https://communitynotes.x.com/guide/en/about/faq>, 2024.
- [297] BuzzFeed News, *Facebook knows that adding labels to Trump's false claims does little to stop their spread*, <https://www.buzzfeednews.com/article/craigsilverman/facebook-labels-trump-lies-do-not-stop-spread>, 2020.

- [298] I. Slaughter, A. Peytavin, J. Ugander, and M. Saveski, 'Community notes reduce engagement with and diffusion of false information online,' *Proceedings of the National Academy of Sciences*, vol. 122, no. 38, e2503413122, 2025.
- [299] X Community Notes, *Downloading data*, <https://communitynotes.x.com/guide/en/under-the-hood/download-data>, [Accessed: 2025-05-20], 2025.
- [300] J. J. Van Bavel, C. E. Robertson, K. Del Rosario, J. Rasmussen, and S. Rathje, 'Social media and morality,' *Annual Review of Psychology*, vol. 75, no. 1, pp. 311–340, 2024.
- [301] Society of Professional Journalists, *Code of ethics*, <https://www.spj.org/ethicscode.asp>, [Accessed: 2024-08-26].
- [302] World Economic Forum, *Global risks report*, <https://www.weforum.org/publications/global-risks-report-2024/>, [Accessed: 2024-08-26].
- [303] C. P. Drolsbach and N. Pröllochs, 'Content moderation on social media in the EU: Insights from the DSA Transparency Database,' in *Companion Proceedings of the ACM on Web Conference 2024*, 2024, pp. 939–942.
- [304] T. Renault, D. R. Amariles, and A. Troussel, 'Collaboratively adding context to social media posts reduces the sharing of false news,' *ArXiv*, 2024.
- [305] J. S. Lerner, Y. Li, P. Valdesolo, and K. S. Kassam, 'Emotion and decision making,' *Annual Review of Psychology*, vol. 66, no. 1, pp. 799–823, 2015.
- [306] H. Zade et al., 'To reply or to quote: Comparing conversational framing strategies on Twitter,' *ACM Journal on Computing and Sustainable Societies*, vol. 2, no. 1, pp. 1–27, 2024.
- [307] F. Gaisbauer, A. Pournaki, S. Banisch, and E. Olbrich, 'Ideological differences in engagement in public debate on Twitter,' *PLoS ONE*, vol. 16, no. 3, e0249241, 2021.
- [308] A. Kozyreva et al., 'Resolving content moderation dilemmas between free speech and harmful misinformation,' *Proceedings of the National Academy of Sciences*, vol. 120, no. 7, e2210666120, 2023.
- [309] D. Neumann and N. Rhodes, 'Morality in social media: A scoping review,' *New Media & Society*, vol. 26, no. 2, pp. 1096–1126, 2024.
- [310] M. J. Crockett, 'Moral outrage in the digital age,' *Nature Human Behaviour*, vol. 1, no. 11, pp. 769–771, 2017.
- [311] J. M. Salerno and L. C. Peter-Hagene, 'The interactive effect of anger and disgust on moral outrage and judgments,' *Psychological Science*, vol. 24, no. 10, pp. 2069–2078, 2013.
- [312] W. J. Brady, K. McLoughlin, T. N. Doan, and M. J. Crockett, 'How social learning amplifies moral outrage expression in online social networks,' *Science Advances*, vol. 7, no. 33, eabe5641, 2021.
- [313] Statista, *Number of internet and social media users worldwide as of july 2024*, <https://www.statista.com/statistics/617136/digital-population-worldwide/>, [Accessed: 2024-08-26].
- [314] W. Brannon and D. Roy, 'The speed of news in Twitter (X) versus radio,' *Scientific Reports*, vol. 14, no. 1, p. 11 939, 2024.
- [315] Y. Sang, J. Y. Lee, S. Park, C. Fisher, and G. Fuller, 'Signalling and expressive interaction: Online news users' different modes of interaction on digital platforms,' *Digital Journalism*, vol. 8, no. 4, pp. 467–485, 2020.
- [316] N. Karusala and R. Anderson, 'Towards conviviality in navigating health information on social media,' in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–14.
- [317] J. Zhou, Y. Zhang, Q. Luo, A. G. Parker, and M. De Choudhury, 'Synthetic lies: Understanding AI-generated misinformation and evaluating algorithmic and human solutions,' in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–20.
- [318] D. Gamage, P. Ghasiya, V. Bonagiri, M. E. Whiting, and K. Sasahara, 'Are deepfakes concerning? Analyzing conversations of deepfakes on Reddit and exploring societal implications,' in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–19.
- [319] Y. Zhang et al., 'What do we mean when we talk about trust in social media? A systematic review,' in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–22.
- [320] R. Dahlke and J. Hancock, 'Effects of misinformation on election beliefs: Disentangling motivated reasoning from selective exposure,' *OSF*, 2024.

- [321] SSRS, *CNN Poll on Biden, economy and elections*, <https://www.documentcloud.org/documents/23895856-cnn-poll-on-biden-economy-and-elections>, [Accessed: 2024-08-26].
- [322] K. L. McLoughlin, W. J. Brady, A. Goolsbee, B. Kaiser, K. Klonick, and M. Crockett, 'Misinformation exploits outrage to spread online,' *Science*, vol. 386, no. 6725, pp. 991–996, 2024.
- [323] T. Wilner, K. Mimizuka, A. Bhimdiwala, J. C. Young, and A. Arif, 'It's about time: Attending to temporality in misinformation interventions,' in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–19.
- [324] M. McClure Haughey, M. Povolo, and K. Starbird, 'Bridging contextual and methodological gaps on the "misinformation beat": Insights from journalist-researcher collaborations at speed,' in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–15.
- [325] H. Tang, G. Lenzini, S. Greiff, B. Rohles, and A. Sergeeva, "'Who knows? Maybe it really works": Analysing users' perceptions of health misinformation on social media,' in *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, 2024, pp. 1499–1517.
- [326] P. Juneja, M. M. Bhuiyan, and T. Mitra, 'Assessing enactment of content regulation policies: A post hoc crowd-sourced audit of election misinformation on YouTube,' in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–22.
- [327] B. Schaffner et al., "'Community guidelines make this the best party on the internet': An in-depth study of online platforms' content moderation policies,' in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–16.
- [328] H. Heuer and E. L. Glassman, 'A comparative evaluation of interventions against misinformation: Augmenting the WHO checklist,' in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–21.
- [329] F. Jahanbakhsh and D. R. Karger, 'A browser extension for in-place signaling and assessment of misinformation,' in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–21.
- [330] A. Hassoun, I. Beacock, S. Consolvo, B. Goldberg, P. G. Kelley, and D. M. Russell, 'Practicing information sensibility: How Gen Z engages with online information,' in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–17.
- [331] S. Stieglitz and L. Dang-Xuan, 'Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior,' *Journal of Management Information Systems*, vol. 29, no. 4, pp. 217–248, 2013.
- [332] J. Berger and K. L. Milkman, 'What makes online content viral?' *Journal of Marketing Research*, vol. 49, no. 2, pp. 192–205, 2012.
- [333] J. Jakubik, M. Vössing, N. Pröllochs, D. Bär, and S. Feuerriegel, 'Online emotions during the storming of the US Capitol: Evidence from the social media network Parler,' in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 17, 2023, pp. 423–434.
- [334] A. Yadollahi, A. G. Shahraki, and O. R. Zaiane, 'Current state of text sentiment analysis from opinion to emotion mining,' *ACM Computing Surveys (CSUR)*, vol. 50, no. 2, pp. 1–33, 2017.
- [335] F. Zollo et al., 'Emotional dynamics in the age of misinformation,' *PLoS ONE*, vol. 10, no. 9, e0138740, 2015.
- [336] M. Del Vicario et al., 'Echo chambers: Emotional contagion and group polarization on Facebook,' *Scientific Reports*, vol. 6, no. 1, p. 37 825, 2016.
- [337] B. Pang, L. Lee, et al., 'Opinion mining and sentiment analysis,' *Foundations and Trends® in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.
- [338] P. Ekman, 'An argument for basic emotions,' *Cognition & Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.
- [339] D. A. Sauter, F. Eisner, P. Ekman, and S. K. Scott, 'Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations,' *Proceedings of the National Academy of Sciences*, vol. 107, no. 6, pp. 2408–2412, 2010.
- [340] P. Ekman, W. V. Friesen, and P. Ellsworth, *Emotion in the human face: Guidelines for research and an integration of findings*. Elsevier, 2013, vol. 11.

- [341] C. Martel, G. Pennycook, and D. G. Rand, 'Reliance on emotion promotes belief in fake news,' *Cognitive Research: Principles and Implications*, vol. 5, pp. 1–20, 2020.
- [342] J. Haidt et al., 'The moral emotions,' *Handbook of Affective Sciences*, vol. 11, no. 2003, pp. 852–870, 2003.
- [343] J. Van de Vyver and D. Abrams, 'Testing the prosocial effectiveness of the prototypical moral emotions: Elevation increases benevolent behaviors and outrage increases justice behaviors,' *Journal of Experimental Social Psychology*, vol. 58, pp. 23–33, 2015.
- [344] W. J. Brady, K. L. McLoughlin, M. P. Torres, K. F. Luo, M. Gendron, and M. Crockett, 'Overperception of moral outrage in online social networks inflates beliefs about intergroup hostility,' *Nature Human Behaviour*, vol. 7, no. 6, pp. 917–927, 2023.
- [345] K. Solovev and N. Pröllochs, 'Moralized language predicts hate speech on social media,' *PNAS Nexus*, vol. 2, no. 1, pgac281, 2023.
- [346] Z. Shah, J. Chu, S. Qaisar, Z. Hassan, and U. Ghani, 'Perceived public condemnation and avoidance intentions: The mediating role of moral outrage,' *Journal of Public Affairs*, vol. 20, no. 1, e2027, 2020.
- [347] K. Rost, L. Stahel, and B. S. Frey, 'Digital social norm enforcement: Online firestorms in social media,' *PLoS ONE*, vol. 11, no. 6, e0155923, 2016.
- [348] W. Strathern, M. Schoenfeld, R. Ghawi, and J. Pfeffer, 'Against the others! Detecting moral outrage in social media networks,' in *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2020, pp. 322–326.
- [349] J. Camacho-collados et al., 'TweetNLP: Cutting-edge natural language processing for social media,' in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2022, pp. 38–49.
- [350] D. Loureiro, F. Barbieri, L. Neves, L. Espinosa Anke, and J. Camacho-collados, 'TimeLMs: Diachronic language models from Twitter,' in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2022, pp. 251–260.
- [351] J. Hartmann, *Emotion english DistilRoBERTa-base*, <https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/>, [Accessed: 2023-09-01].
- [352] S. Butt, S. Sharma, R. Sharma, G. Sidorov, and A. Gelbukh, 'What goes on inside rumour and non-rumour tweets and their reactions: A psycholinguistic analyses,' *Computers in Human Behavior*, vol. 135, p. 107345, 2022.
- [353] D. Rozado, R. Hughes, and J. Halberstadt, 'Longitudinal analysis of sentiment and emotion in news media headlines using automated labelling with transformer language models,' *PLoS ONE*, vol. 17, no. 10, e0276367, 2022.
- [354] G. W. Imbens and T. Lemieux, 'Regression discontinuity designs: A guide to practice,' *Journal of Econometrics*, vol. 142, no. 2, pp. 615–635, 2008.
- [355] M. D. Cattaneo and R. Titiunik, 'Regression discontinuity designs,' *Annual Review of Economics*, vol. 14, no. 1, pp. 821–851, 2022.
- [356] A. Goldenberg and J. J. Gross, 'Digital emotion contagion,' *Trends in Cognitive Sciences*, vol. 24, no. 4, pp. 316–328, 2020.
- [357] A. D. I. Kramer, J. E. Guillory, and J. T. Hancock, 'Experimental evidence of massive-scale emotional contagion through social networks,' *Proceedings of the National Academy of Sciences*, vol. 111, no. 24, pp. 8788–8790, 2014.
- [358] H. Rosenbusch, A. M. Evans, and M. Zeelenberg, 'Multilevel emotion transfer on youtube: Disentangling the effects of emotional contagion and homophily on video audiences,' *Social Psychological and Personality Science*, vol. 10, no. 8, pp. 1028–1035, 2019.
- [359] C. Hausman and D. S. Rapson, 'Regression discontinuity in time: Considerations for empirical applications,' *Annual Review of Resource Economics*, vol. 10, no. 1, pp. 533–552, 2018.
- [360] L. Keele and N. J. Kelly, 'Dynamic models for dynamic theories: The ins and outs of lagged dependent variables,' *Political Analysis*, vol. 14, no. 2, pp. 186–205, 2006.
- [361] X, *About replies and mentions*, <https://help.x.com/en/using-x/mentions-and-replies>, [Accessed: 2024-08-26].

- [362] L. M. Smith, L. E. Lévesque, J. S. Kaufman, and E. C. Strumpf, 'Strategies for evaluating the assumptions of the regression discontinuity design: A case study using a human papillomavirus vaccination programme,' *International Journal of Epidemiology*, vol. 46, no. 3, pp. 939–949, 2017.
- [363] T. Lumley, P. Diehr, S. Emerson, and L. Chen, 'The importance of the normality assumption in large public health data sets,' *Annual Review of Public Health*, vol. 23, no. 1, pp. 151–169, 2002.
- [364] M. W. Fagerland, 'T-tests, non-parametric tests, and large studies—a paradox of statistical practice?' *BMC Medical Research Methodology*, vol. 12, pp. 1–7, 2012.
- [365] J. Kim, C. Jeong, S. Park, M. Cha, and W. Lee, 'How do moral emotions shape political participation? A cross-cultural analysis of online petitions using language models,' in *Findings of the Association for Computational Linguistics ACL 2024*, 2024, pp. 16 274–16 289.
- [366] C. Hutto and E. Gilbert, 'VADER: A parsimonious rule-based model for sentiment analysis of social media text,' in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, 2014, pp. 216–225.
- [367] YouTube, *Testing new ways to offer viewers more context and information on videos*, <https://blog.youtube/news-and-events/new-ways-to-offer-viewers-more-context/>, [Accessed: 2025-05-25].
- [368] C. Molho, J. M. Tybur, E. Güler, D. Balliet, and W. Hofmann, 'Disgust and anger relate to different aggressive responses to moral violations,' *Psychological Science*, vol. 28, no. 5, pp. 609–619, 2017.
- [369] F. van der Eijk and S. Columbus, 'Expressions of moral disgust reflect both disgust and anger,' *Cognition and Emotion*, vol. 37, no. 3, pp. 499–514, 2023.
- [370] S. Wijenayake, D. Hettiachchi, S. Hosio, V. Kostakos, and J. Goncalves, 'Effect of conformity on perceived trustworthiness of news in social media,' *IEEE Internet Computing*, vol. 25, no. 1, pp. 12–19, 2020.
- [371] J. 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*, vol. 97, pp. 202–215, 2019.
- [372] K. Peng, Y. Zheng, Y. Qiu, and Q. Li, 'Rage of righteousness: Anger's role in promoting debunking effects during the COVID-19 pandemic,' *Computers in Human Behavior*, vol. 148, p. 107 896, 2023.
- [373] M. Mosleh, C. Martel, D. Eckles, and D. Rand, 'Perverse downstream consequences of debunking: Being corrected by another user for posting false political news increases subsequent sharing of low quality, partisan, and toxic content in a Twitter field experiment,' in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–13.
- [374] A. Efstratiou and E. De Cristofaro, 'Adherence to misinformation on social media through socio-cognitive and group-based processes,' *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW2, pp. 1–35, 2022.
- [375] S. Mohammadi et al., 'From Birdwatch to Community Notes, from Twitter to X: Four years of community-based content moderation,' *ArXiv*, 2025.
- [376] B. T. Truong, S. Wu, A. Flammini, F. Menczer, and A. J. Stewart, 'Community notes are vulnerable to rater bias and manipulation,' *ArXiv*, 2025.
- [377] P. Bouchaud and P. Ramaciotti, 'Algorithmic resolution of crowd-sourced moderation on X in polarized settings across countries,' *ArXiv*, 2025.
- [378] T. Renault, M. Mosleh, and D. G. Rand, 'Republicans are flagged more often than Democrats for sharing misinformation on X's Community Notes,' *Proceedings of the National Academy of Sciences*, vol. 122, no. 25, e2502053122, 2025.
- [379] M. Luca and G. Zervas, 'Fake it till you make it: Reputation, competition, and Yelp review fraud,' *Management Science*, vol. 62, no. 12, pp. 3412–3427, 2016.
- [380] G. Pennycook and D. G. Rand, 'Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning,' *Cognition*, vol. 188, pp. 39–50, 2019.
- [381] D. M. Kahan, E. Peters, E. C. Dawson, and P. Slovic, 'Motivated numeracy and enlightened self-government,' *Behavioural Public Policy*, vol. 1, no. 1, pp. 54–86, 2017.
- [382] M. D. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini, 'Political polarization on Twitter,' in *International AAAI Conference on Web and Social Media*, 2011.

- [383] P. Barberá, J. T. Jost, J. Nagler, J. A. Tucker, and R. Bonneau, 'Tweeting from left to right: Is online political communication more than an echo chamber?' *Psychological Science*, vol. 26, no. 10, pp. 1531–1542, 2015.
- [384] K. Solovev and N. Pröllochs, 'Hate speech in the political discourse on social media: Disparities across parties, gender, and ethnicity,' in *Proceedings of the ACM web conference 2022*, 2022.
- [385] K. Toyoda et al., 'Understanding and mitigating polarization in community notes: Factors and strategies for improved consensus,' in *Companion Proceedings of the ACM on Web Conference 2025*, 2025, pp. 2694–2698.
- [386] N. Borenstein, G. Warren, D. Elliott, and I. Augenstein, 'Can community notes replace professional fact-checkers?' In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2025, pp. 535–552.
- [387] X Community Notes, *Note selection algorithm*, <https://github.com/twitter/communitynotes/tree/91b24d1d346e43dfd3b196f7fe6af5e2e8619094>, 2024.
- [388] Z. Arjmandi-Lari, A. Mantzarlis, and T. Stafford, 'Threats to the sustainability of Community Notes on X,' *ArXiv*, 2025.
- [389] O. Razuvayevskaya, A. Tayebi, U. D. Sørensen, K. Bontcheva, and R. Rogers, 'Timeliness, consensus, and composition of the crowd: Community Notes on X,' *ArXiv*, 2025.
- [390] B. He, Y. Hu, Y.-C. Lee, S. Oh, G. Verma, and S. Kumar, 'A survey on the role of crowds in combating online misinformation: Annotators, evaluators, and creators,' *ACM Transactions on Knowledge Discovery from Data*, vol. 19, no. 1, pp. 1–30, 2025.
- [391] O. Vinhas and M. Bastos, 'Fact-checking misinformation: Eight notes on consensus reality,' *Journalism Studies*, vol. 23, no. 4, pp. 448–468, 2022.
- [392] N. Walter, J. Cohen, R. L. Holbert, and Y. Morag, 'Fact-checking: A meta-analysis of what works and for whom,' *Political Communication*, vol. 37, no. 3, pp. 350–375, 2020.
- [393] S. Kankham and J.-R. 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*, vol. 41, no. 12, pp. 7711–7725, 2025.
- [394] C. Jia, A. Y. Lee, R. C. Moore, C. H.-S. Decatur, S. X. Liu, and J. T. Hancock, 'Collaboration, crowdsourcing, and misinformation,' *PNAS nexus*, vol. 3, no. 10, p. pgae434, 2024.
- [395] X Community Notes, *Agreement safeguards*, <https://communitynotes.x.com/guide/en/under-the-hood/ranking-notes#agreement-safeguards>, 2025.
- [396] H. Li et al., 'Scaling human judgment in community notes with LLMs,' *ArXiv*, 2025.
- [397] N. Hassan, M. Yousuf, M. Mahfuzul Haque, J. A. Suarez Rivas, and M. Khadimul Islam, 'Examining the roles of automation, crowds and professionals towards sustainable fact-checking,' in *Companion Proceedings of The 2019 World Wide Web Conference*, 2019, pp. 1001–1006.
- [398] S. De, M. A. Bakker, J. Baxter, and M. Saveski, 'Supernotes: Driving consensus in crowd-sourced fact-checking,' in *Proceedings of the ACM on Web Conference 2025*, 2025, pp. 3751–3761.
- [399] J. Wu, Z. Fu, H. Wang, F. Li, and M.-Y. Kan, 'Beyond the crowd: LLM-augmented community notes for governing health misinformation,' *ArXiv*, 2025.
- [400] S. Mohammadi and T. Yasseri, 'AI feedback enhances community-based content moderation through engagement with counterarguments,' *ArXiv*, 2025.
- [401] G. Pennycook and D. G. Rand, 'The psychology of fake news,' *Trends in Cognitive Sciences*, vol. 25, no. 5, pp. 388–402, 2021.
- [402] J. Allen, D. J. Watts, and D. G. Rand, 'Quantifying the impact of misinformation and vaccine-skeptical content on facebook,' *Science*, vol. 384, no. 6699, p. eadk3451, 2024.
- [403] A. Presser, *Testing a new feature to enhance content on TikTok*, <https://newsroom.tiktok.com/en-us/footnotes>, [Accessed: 2025-05-25].
- [404] J. Kaplan, *More speech and fewer mistakes*, <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>, [Accessed: 2025-05-25].
- [405] X Community Notes, *Request a note*, <https://communitynotes.x.com/guide/en/under-the-hood/note-requests>, [Accessed: 2025-05-20].

- [406] C. Martel and D. G. Rand, 'Fact-checker warning labels are effective even for those who distrust fact-checkers,' *Nature Human Behaviour*, vol. 8, no. 10, pp. 1957–1967, 2024.
- [407] J. Ni, M. Shi, D. Stambach, M. Sachan, E. Ash, and M. Leippold, 'AFaCTA: Assisting the annotation of factual claim detection with reliable LLM annotators,' pp. 1890–1912, 2024.
- [408] C.-H. Chiang and H.-y. Lee, 'Can large language models be an alternative to human evaluations?' In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 15 607–15 631.
- [409] P. Pataranutaporn, C. Archiwanguprok, S. W. Chan, E. Loftus, and P. Maes, 'Synthetic human memories: AI-edited images and videos can implant false memories and distort recollection,' in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025, pp. 1–20.
- [410] X Community Notes, *Media matching*, <https://communitynotes.x.com/guide/en/under-the-hood/media-matching>, [Accessed: 2025-09-09].
- [411] H. Lin et al., 'High level of correspondence across different news domain quality rating sets,' *PNAS Nexus*, vol. 2, no. 9, pgad286, 2023.
- [412] S. F. Kuuse, U. Kangur, R. Chakraborty, and R. Sharma, 'Crowdsourced fact-checking or biased commentary? Analyzing political bias in Twitter's Community Notes,' in *Companion Proceedings of the ACM on Web Conference 2025*, 2025, pp. 2661–2669.
- [413] X Community Notes, *Top contributors*, <https://communitynotes.x.com/guide/en/contributing/top-contributors>, [Accessed: 2025-05-20].
- [414] X Community Notes, *Open-source code*, <https://communitynotes.x.com/guide/en/under-the-hood/note-ranking-code>, [Accessed: 2025-05-20].
- [415] C. Martel, J. Allen, G. Pennycook, and D. G. Rand, 'Political motives help rather than hinder crowdsourced fact-checking,' *OSF*, 2025.
- [416] X Community Notes, *AI note writers*, <https://communitynotes.x.com/guide/en/api/overview>, [Accessed: 2025-09-09].
- [417] I. Augenstein et al., 'Factuality challenges in the era of large language models and opportunities for fact-checking,' *Nature Machine Intelligence*, vol. 6, no. 8, pp. 852–863, 2024.
- [418] X Community Notes, *Note ranking algorithm*, <https://communitynotes.twitter.com/guide/en/under-the-hood/ranking-notes>, 2024.
- [419] X. Zhang et al., 'TwHIN-BERT: A socially-enriched pre-trained language model for multilingual tweet representations at Twitter,' *ArXiv*, 2022.
- [420] T. Wolf et al., 'Transformers: State-of-the-art natural language processing,' *ACL Anthology*, pp. 38–45, Oct. 2020.
- [421] S. Guo and M. W. Fraser, *Propensity score analysis: Statistical methods and applications*. SAGE Publications, 2014, vol. 11.
- [422] H. Dette and M. Schumann, 'Testing for equivalence of pre-trends in difference-in-differences estimation,' *Journal of Business & Economic Statistics*, vol. 42, no. 4, pp. 1289–1301, 2024.
- [423] D. Lakens, 'Equivalence tests: A practical primer for t tests, correlations, and meta-analyses,' *Social Psychological and Personality Science*, vol. 8, no. 4, pp. 355–362, 2017.