

Essays on Understanding and Combating Misinformation at Scale

by

Jennifer Allen

B.A. Computer Science and Psychology, Yale, 2016
S.M. Management Research, MIT Sloan School of Management, 2022

Submitted to the Department of Management
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN MANAGEMENT

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2024

© 2024 Jennifer Allen. This work is licensed under a [CC BY-NC-ND 4.0](#) license.

The author hereby grants to MIT a nonexclusive, worldwide, irrevocable, royalty-free license to exercise any and all rights under copyright, including to reproduce, preserve, distribute and publicly display copies of the thesis, or release the thesis under an open-access license.

Authored by: Jennifer Allen
Department of Management
May 3, 2024

Certified by: David G. Rand
Department of Management
Thesis Supervisor

Accepted by: Eric So
Professor, Global Economics and Finance
Faculty Chair, MIT Sloan PhD Program

Essays on Understanding and Combating Misinformation at Scale

by

Jennifer Allen

Submitted to the Department of Management
on May 3, 2024 in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN MANAGEMENT

ABSTRACT

In Chapter 1, I explore the use of crowdsourcing as a potential solution to the misinformation problem at scale. Perhaps the most prominent approach to combating misinformation is the use of professional fact-checkers. This approach, however, is not scalable: Professional fact-checkers cannot possibly keep up with the volume of misinformation produced every day. Furthermore, many people see fact-checkers as having a liberal bias and thus distrust them. Here, we explore a potential solution to both of these problems: leveraging the “wisdom of crowds” to make fact-checking possible at scale using politically-balanced groups of laypeople. Our results indicate that crowdsourcing is a promising approach for helping to identify misinformation at scale.

In Chapter 2, joint with David Rand and Cameron Martel, I extend work on crowdsourced fact-checking to assess the viability of crowdsourcing in an opt-in, polarized environment. We leverage data from Birdwatch, Twitter’s crowdsourced fact-checking pilot program, to examine how shared partisanship affects participation in crowdsourced fact-checking. Our findings provide clear evidence that Birdwatch users preferentially challenge content from those with whom they disagree politically. While not necessarily indicating that Birdwatch is ineffective for identifying misleading content, these results demonstrate the important role that partisanship can play in content evaluation. Platform designers must consider the ramifications of partisanship when implementing crowdsourcing programs.

In Chapter 3, I examine the role of online (mis)information on US vaccine hesitancy. I combine survey experimental estimates of persuasion with exposure data from Facebook to estimate the extent to which (mis)information content on Facebook reduces COVID vaccine acceptance. Contrary to popular belief, I find that factually-accurate vaccine-skeptical content was approximately 50X more impactful than outright false misinformation. Although outright misinformation had a larger negative effect per exposure on vaccination intentions than factually accurate content, it was rarely seen on social media. In contrast, mainstream media articles reporting on rare deaths following vaccination garnered hundreds of millions of views. While this work suggests that limiting the spread of misinformation has important public health benefits, it highlights the need to scrutinize accurate-but-misleading content published by mainstream sources.

Thesis supervisor: David G. Rand

Title: Erwin H. Schell Professor, Department of Management

Acknowledgments

This work would not have been possible without the support of my mentors, collaborators, friends, and family. First, my committee – David Rand, Abdullah Almaatouq, and Adam Berinsky, who have been amazing mentors throughout my time at MIT. I am particularly grateful to have had Dave Rand as my advisor – a person who is not only is a brilliant researcher but an incredible mentor and advocate. I’m so glad our taste in research overlaps more than our taste in music (with the exception of Olivia Rodrigo).

Beyond his many research talents, I am also incredibly appreciate of Dave’s ability to build an amazing community in the Human Cooperation Lab. In particular, I want to thank Adam, Luke, and Ben for responding to all my frantic stats emails; Nick, for banging the drum on attrition; Cameron, for always having my back on research (and for back-channeling when I needed it most); Gord, for keeping me on my toes; Hause, Brian, Tom, and Ashley for being post-doc extraordinaires, and all other members of the HCL that have made up my research and social world over this PhD journey.

I’ve had many other mentors along the way. I would not be here in any way without the Computational Social Science Group at Microsoft Research – in particular, Duncan Watts and David Rothschild– who introduced me to the Computational Social Science community and started me on this PhD journey. I owe so much of who am I as a researcher to their mentorship, and I’m so grateful that I took the leap to academia with their guidance. I owe another thank you to the ACRONYM crew – Sol Messing, Alex Coppock, Minali Aggarwal, Dan Frankowski – thank you for opening my eyes to the crazy world of online political ads (although I don’t know if I was ready for what was on the other side).

Finally, I want to acknowledge my family and friends. I could not have survived without the loving support of my friends – Catherine, Gina Starfield, Emma, Lara, Truett, Devon, Steph, Sophie, Sukhi, Sylvan, Alex, Olivia and many many more, who kept me sane over the many ups and downs over the last 5 years. I thank my partner Reed, who had to put up with countless practice job talks, email pitches, and panicked phone calls, and who has been my rock through this process. Lastly, I want to thank my family – my mom Chinhui, my dad Eddie, and my sister Jessica. I’m so grateful for their emotional support throughout the years. Thank you for tolerating me at my most frazzled state.

This dissertation is dedicated to my mom – as a (very inadequate) thank you for the love, encouragement, and inspiration she has shown me throughout my life. She is my personal and academic role model, and I couldn’t have done it without her.

Contents

Title page	1
Abstract	3
Acknowledgments	5
List of Figures	11
List of Tables	13
1 Scaling Up Fact-checking Using the Wisdom of Crowds	15
1.1 Introduction	15
1.2 Materials and Methods	16
1.2.1 Materials	17
1.2.2 Methods	17
1.2.3 Fact-checkers	17
1.2.4 Participants	18
1.2.5 Analysis	18
1.3 Results	19
1.3.1 Fact-checkers	19
1.3.2 Correlation between Fact-Checkers and the Crowd	19
1.3.3 AUROC Analysis	20
1.3.4 Comparing Crowds of Different Compositions	21
1.4 Discussion	22
2 Birds of a Feather Don't Fact-check Each Other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program	27
2.1 Introduction	27
2.1.1 The Birdwatch Platform	28
2.1.2 The Current Research	29
2.2 Related Work	29
2.2.1 Partisanship and Online Behavior	29
2.2.2 Motivated Reasoning	30
2.2.3 Crowdsourced Fact-checking	31
2.3 Methods	31

2.3.1	Twitter Datasets	31
2.3.2	Features	32
2.3.3	Models	34
2.4	Results	36
2.4.1	Predicting Misleadingness Classification	36
2.4.2	Helpfulness Classification Results	37
2.4.3	Shared partisanship predicts classifications and ratings	39
2.5	Discussion	41
3	Quantifying the Impact of Misinformation and Vaccine-Skeptical Content on Facebook	51
3.1	Introduction	51
3.2	Methods	54
3.2.1	Facebook Exposure Data	54
3.2.2	Survey Experiments	56
3.2.3	Facebook URL Predicted Treatment Effects	60
3.2.4	Predicting Treatment Effects for Facebook URLs	62
3.2.5	Combining Predicted Treatment Effects and Facebook Viewership Data	63
3.3	Results	63
3.3.1	Survey Experiments	63
3.3.2	Exposure on Facebook	67
3.3.3	Predicting Treatment Effects for Facebook URLs	68
3.3.4	Impact Estimates	70
3.4	Discussion	73
A	Chapter 2 Appendix	79
A.1	Model Robustness	79
A.1.1	Alternate Evaluation Metrics	79
A.2	Logistic Regression, Full Results	81
B	Chapter 3 Appendix	83
B.0.1	Study Variable Definitions	83
B.0.2	Model	84
B.0.3	Balance Checks	85
B.0.4	Differential Attrition	86
B.0.5	Additional Survey Results	88
B.0.6	Subject Level Heterogeneity	88
B.0.7	Cutoff Tuning	90
B.0.8	Top Viewed URLs	91
B.0.9	Most Harmful Domains	91
B.0.10	Contemporaneous Treatment Effect Estimates	94
B.0.11	Formal Model	96
B.0.12	Alternative Models	97
B.0.13	Impact Calculation	97
B.0.14	Confidence Intervals for Impact Estimates	99

B.0.15 Threshold for Skeptical URLs	99
B.0.16 Facebook Subject-Level Heterogeneity	99
B.0.17 Crowdsourcing Variable Definition	100
B.0.18 Crowdsourcing Performance	101

References	105
-------------------	------------

List of Figures

1.1	Wisdom of Crowds Correlation Analysis	20
1.2	Wisdom of Crowds AUROC Analysis	24
1.3	Comparing Crowd Performance	25
2.1	Example Birdwatch Tweet and Note	45
2.2	Histogram Birdwatch Notes	46
2.3	Histogram Birdwatch Ratings	47
2.4	Predicting Misleadingness	47
2.5	Predicting Helpfulness	48
2.6	Misleading Classifications by Partisanship	48
2.7	Helpful Ratings by Partisanship	49
2.8	Helpfulness Score by Percent Co-partisans	49
3.1	High-Level Overview of Treatment Effect Prediction	60
3.2	Effect of Misinformation on Vaccination Intentions	65
3.3	Moderators Vaccine Intentions	66
3.4	Harmful to Health on Vaccine Intetions	67
3.5	Exposure to Vaccine Content on Facebook	69
3.6	Crowd Judgments Predict Treatment Effects	71
3.7	Distribution of Facebook Predicted Treatment Effects	72
3.8	Impact of Facebook on Vaccination Intentions	73
3.9	Impact of Facebook on Vaccination Intentions, Low vs. High Quality Domains	74
A.1	Compare RF Models, Misleading	80
A.2	Compare RF Models, Helpfulness	80
B.1	Moderators Study 1 and 2	88
B.2	Causal Forest	90
B.3	Cutoff Tuning	92
B.4	Top Viewed URLs Facebook	93
B.5	Top Harmful Domains, Weighted by Views	93
B.6	Top harmful / hesitancy-inducing domains	94
B.7	Simulate Promoting Vaccine Content	95
B.8	Varying cutoffs impact on results	100
B.9	Percent “vaccine-skeptical” content by various demographics groups	101

List of Tables

2.1	Content Features Birdwatch	33
2.2	Context Features Birdwatch	34
2.3	Descriptive Statistics Birdwatch	35
2.4	Feature Sets Misleadingness	36
2.5	Feature Sets Helpfulness	38
2.6	Notes: Misleading by Party	40
2.7	Ratings: Helpful by Party	40
3.1	Labeled URL Data	61
3.2	Performance for a model predicting the crowdsourced aggregate score of the Facebook URL	63
3.3	Performance for a model predicting the binary skeptical vs. promoting/not skeptical rating of the Facebook URL	63
A.1	Logistic Regression Misleadingness	81
A.2	Logistic Regression Helpfulness	82
B.1	Balance Check, Study 1	86
B.2	Balance Check, Study 2	86
B.3	Attrition Check, Study 1	87
B.4	Attrition Check, Study 2	87
B.5	Study 1: Individual Level Heterogeneity	89
B.6	Study 2: Individual Level Heterogeneity	89
B.7	P-values from omnibus test of heterogeneity from causal forest models	90
B.8	Performance Metrics for Alternate Methods	97
B.9	Performance Metrics for Alternate Methods, Binary Classification Task	98
B.10	Model Comparison, Crowdsourcing	103

Chapter 1

Scaling Up Fact-checking Using the Wisdom of Crowds

1.1 Introduction

With concerns about fake news growing in the leadup to the 2020 U.S. presidential election, many people have questioned social media platforms’ ability to combat disinformation. In response, Facebook, Twitter, and Google have invested in fact-checking as a way to combat misconception [1]–[3]. However, although many studies have shown that fact-checking can be effective in correcting misconceptions [4], the strategy has problems with both scalability and trust [5].

Fact-checking is a laborious process that cannot keep pace with the enormous amount of content on social media. For example, according to a recent article published by The Hill, in 2020, Facebook’s six fact-checking partners had a combined 26 full-time staff that fact-check roughly 200 pieces of content per month – a tiny fraction of potentially inaccurate content on Facebook [6]. Furthermore, according to a Poynter study, 50% of Americans (and 70% of Republicans) believe that fact-checkers are biased and distrust fact-checking corrections [7]. Our study explores a solution to these problems of credibility and scale: Applying the “wisdom of crowds” to fact-checking.

Pennycook and Rand [8] previously showed that laypeople across the political spectrum are surprisingly good at distinguishing high from low quality sources. Here we ask how well laypeople can tackle the substantially harder problem of rating the veracity of individual articles.

Despite the previous success of crowdsourcing source-level ratings, it is a priori unclear whether the crowd’s ability to discern false from true news extends from sources to headlines. While prior results by the authors show that the crowd’s trust ratings correlate very strongly with experts’ ratings, they also find that familiarity is a key driver of trust [9]. Laypeople tend to distrust sources that are unfamiliar to them, regardless of their journalistic credibility. Since research shows that most people do not consume much news, mainstream or fake, online familiarity is unlikely to be as powerful of a mechanism for discerning true from false headlines [10]–[12].

Indeed, prior research on the ability of individuals to identify fake news has been mixed.

Several studies have shown that individuals consistently rate fake headlines as less plausible than true ones [13]–[15]. However, recent reports have also suggested that ordinary people cannot easily detect false information [16]. One explanation for these conflicting accounts is that the stimuli differ across experiments. For example, asking laypeople to compare an article from The Washington Post to one from a conspiracy site like InfoWars is likely to be an easier task than asking laypeople to discern which of two hyperpartisan Breitbart articles is true.

Our work seeks to address this prior work in two ways. First, in an attempt to produce results as ecologically valid as possible, we sought out a stimulus set that was non-trivially challenging. As part of a collaboration with Facebook, our team was granted access to a set of articles that were flagged as potentially problematic by Facebook’s algorithms. While our goal in this research is to evaluate whether and how platforms like Facebook might implement crowdsourcing to combat fake news, Facebook did not have any role in determining the specific direction or publication of our research. That said, that these URLs are representative of the sort that are currently being sent to third-party fact-checkers is a strong signal of the external validity of our work and how it might be relevant to Facebook’s use-case.

Second, instead of focusing on the individual ability of laypeople to distinguish true and false headlines, we reframe our question in terms of the aggregate performance of the crowd. The wisdom of the crowd, in which the judgment of a diverse, independent group of laypeople outperforms the judgment of a single expert, is a persistent phenomenon across a variety of domains including guessing tasks, medical diagnoses, and corporate earnings [17]–[19]. The literature shows that poor performance at an individual level does not prevent great performance at the aggregate level. A crowd of a sufficient size amplifies the signals of experts and reduces the noise of the uninformed.

Our primary research question examines how well layperson judgments correlate with the judgments of professional fact-checkers and how large of a “crowd” of laypeople is required to achieve reasonable results. As a secondary question, we explore how the inclusion of source information about the domain of the URL (e.g. `breitbart.com`) affects the crowd’s performance. Previous research suggests adding source information could improve laypeople’s ability to detect false news when there is a mismatch between headline plausibility and source trustworthiness. Thus, in our set of headlines that are for the most part relatively implausible (since they are headlines that were flagged by Facebook’s algorithm as potentially false), learning that the headline is from a trusted source could improve discernment.

1.2 Materials and Methods

Data and materials are available online. Participants provided informed consent, and our studies were approved by MIT’s Committee On the Use of Humans as Experimental Subjects, protocol number 1806400195.

1.2.1 Materials

As part of our collaboration, Facebook granted us access to a set of 721 articles that were flagged as potentially inaccurate by their internal algorithm. These articles are a sample of those that Facebook sends to their third-party fact-checking partners. Since research shows that users do not click on most articles they view in their social media feeds [20], we filtered to articles that contained a claim of fact in their headline or lede as determined by four research assistants. We also excluded all broken or removed URLs. This filtering resulted in a subset of 463 articles, of which we randomly selected 225 as primary materials, and left the rest for out-of-sample testing.

1.2.2 Methods

We designed two surveys for each of our target audiences: lay people and professional fact-checkers. On both, participants were asked to assess the central claims of a set of the articles described above. Based on their assessment, participants first determined whether a given article was either: true, misleading, false, or couldn't be determined. Since prior research has shown that asking similar questions multiple times leads to more accurate answers ("the wisdom of many in one mind"), participants also rated each of the articles using 7 questions related to the accuracy and bias of the central claim (7-point Likert scales) [21]. We asked whether the article 1) described an event that actually happened, 2) was true, 3) was accurate 4) was reliable, 5) was trustworthy, 6) was objective, and 7) was written in an unbiased way.

Fact-checkers were presented with the actual URL of each article and asked to research them and to provide any relevant evidence that justified their assessment. Laypeople, on the other hand, were only provided with minimal information to base their assessment. They were not asked to do any research or provide a source for their claims, but rather to rely on their own judgment.

In the interest of finding whether knowledge of the headline's source influenced assessment, one half of the lay participants were randomly assigned to a condition that only displayed the headline and lede of the articles, whereas the other half saw the source domain of the article too (e.g. breitbart.com). To conclude the study and upon completion of the primary task, laypeople were asked to answer the CRT [22] and a series of demographic and political questions.

1.2.3 Fact-checkers

Between 10/27/2019 and 1/21/2020, we recruited three separate fact-checkers from the freelancing site Upwork who had prior experience in fact-checking after an extensive vetting process. First, we identified an initial pool of 20 candidates who all listed fact-checking as one of the skills they offered, and had familiarity with American politics. We then hired three people from this pool to complete an initial assessment task, in which we had them fact-check 20 articles from our set. We then checked their responses to confirm that they were thorough and displayed a mastery of the task, including giving individualized feedback and engaging in discussion when there was substantial disagreement between the fact-checkers. Interestingly, this discussion revealed real, reasoned disagreements, rather than misunderstandings

or sloppiness. Once this initial trial was completed satisfactorily, we had the fact-checkers evaluate the remainder of the articles.

1.2.4 Participants

Between 2/9/2020 and 2/11/2020, we recruited 1,204 US residents from Amazon Mechanical Turk to rate a set of 20 articles each (Mage = 35.15; 38.42% female). On average, each article was rated by 104 participants. With regards to the condition faced by participants, 597 only saw headlines and lede (Mage = 35.19; 39.57% female) and 607 saw the source as well (Mage = 35.11; 37.30% female).

1.2.5 Analysis

Our main analysis used the bootstrap procedure described below to compute the Pearson correlation between the average aggregate accuracy ratings of a politically-balanced crowd and the average fact-checker ratings. We then compared this correlation to a benchmark of the average of the pairwise correlations between the fact-checkers ($r = .62$). We determined the minimum value of n for which the inter-factchecker correlation is (a) included in the 95% CI of the layperson-fact-checker correlation, and (b) is below that 95% CI. To do this bootstrap analysis, we first averaged all participants' ratings across all 7 Likert-scale questions to create an aggregate accuracy score for each person-article pair. A prior factor analysis showed that these questions were highly related and explained by a single factor, so the aggregation served to reduce noise in the individual judgments. We also dichotomized participants into "Democrats" and "Republicans" by asking them to rate their political leaning on a 6-point scale from "Strong Democrat" to "Strong Republican". Then, for each question, we sampled an equal number ($n/2$) of Democrats and Republicans and averaged their responses together. We then computed the correlation across all articles between this politically balanced layperson average and the average of the fact-checkers' aggregated accuracy ratings. We repeated this process 1000 times per article for $n = 2$ to $n = 26$, where n is the total number of laypeople in the crowd. For each n , we reported the averaged Pearson correlation as well as the 95% confidence interval as generated by this bootstrap procedure. We then compared the average ratings and confidence intervals to the average of the pairwise correlations between each of the fact-checkers for the source and no source condition.

While we chose to use Pearson correlation due to its suitability to the task and its familiarity to a general audience, we also find qualitatively similar results for other measures of inter-rater reliability like intra-class correlation.

In addition to the Pearson Correlation, we also performed the same bootstrapping procedure using area-under-the-receiver-operating-curve (AUROC) as an outcome measure. We created labels for each article by first turning the fact-checkers categorical ratings into a binary variable where responses were coded as 1 if the fact-checker labeled the item as "True" and 0 otherwise and then taking the modal fact-checker rating for each article. We used the same bootstrap procedure as above to find the average AUROC and 95% confidence interval for a politically balanced crowd of size 2 to 26 for both the source and no source condition. Additionally, we showed the AUROC curve for a crowd of size 26 to evaluate the trade-off between false-positive and false-negatives for different rating thresholds.

1.3 Results

1.3.1 Fact-checkers

First, as a benchmark, we consider the question of how well the responses of professional fact-checkers correlated with each other. The average correlation between the three professional fact-checkers was .62 (range = .52 - .81, $p < .001$). While this might be considered a “large” correlation according to common social science benchmarks,[23] we consider this level of correlation relatively low for our task considering that fact-checkers rated identical stimuli. Indeed, other measures of inter-rater reliability found only a “fair” or “moderate” level of agreement amongst the fact-checkers.

While perhaps surprising, this result is not anomalous when compared to past research measuring agreement between professional fact-checkers. Other research groups investigating the same phenomenon found similar levels of average correlation among professional fact-checkers [24]. Nor are the results particularly surprising considering the difficulty of the task; the articles provided to us by Facebook had already been flagged by their algorithm as being potentially misleading and thus likely presented a more challenging problem than simply fact-checking a random selection of news. Given that the performance we found was in line with prior work, we used the average inter-factchecker correlation as a benchmark to measure the performance of the crowd.

1.3.2 Correlation between Fact-Checkers and the Crowd

We now turn to consider the performance of the crowd. In our first analysis, we compare the layperson continuous 1 - 7 accuracy ratings to the continuous accuracy ratings of the fact-checkers. We estimate the performance of crowds of different sizes by doing bootstrap-style simulations in which the specified number of Democrat and Republican ratings are drawn with replacement from the full set of ratings for each article. The results of our analysis are found in Figure 1.1

Encouragingly, we find that after approximately 10 politically balanced responses, the crowd is able to match the performance of the fact-checkers. After 8 and 12 responses for the source and no source conditions, respectively, we find that the correlation between laypeople and the fact-checkers does not significantly differ from the average correlation between the fact-checkers (source condition: $n = 8$, $r = .57$, 95% CI = .50 - .64, no source condition: $r = .56$, 95% CI = .51 - .62). After 24 responses, in the source condition, the correlation between the crowd and the fact-checkers is significantly higher than the inter-fact-checker correlation ($n = 22$, $r = .66$, 95% CI = .63 - .70). These results suggest that a crowd of laypeople can correspond better with the average fact-checker than individual fact-checkers correspond with each other.

We also find evidence that supports including source information in our labeling task. We find that the correlations in the source condition are significantly higher than in the no source condition ($n = 26$, $p < .001$). This finding supports our hypothesis that including source information adds a valuable signal in cases where the accuracy of the headline is ambiguous.

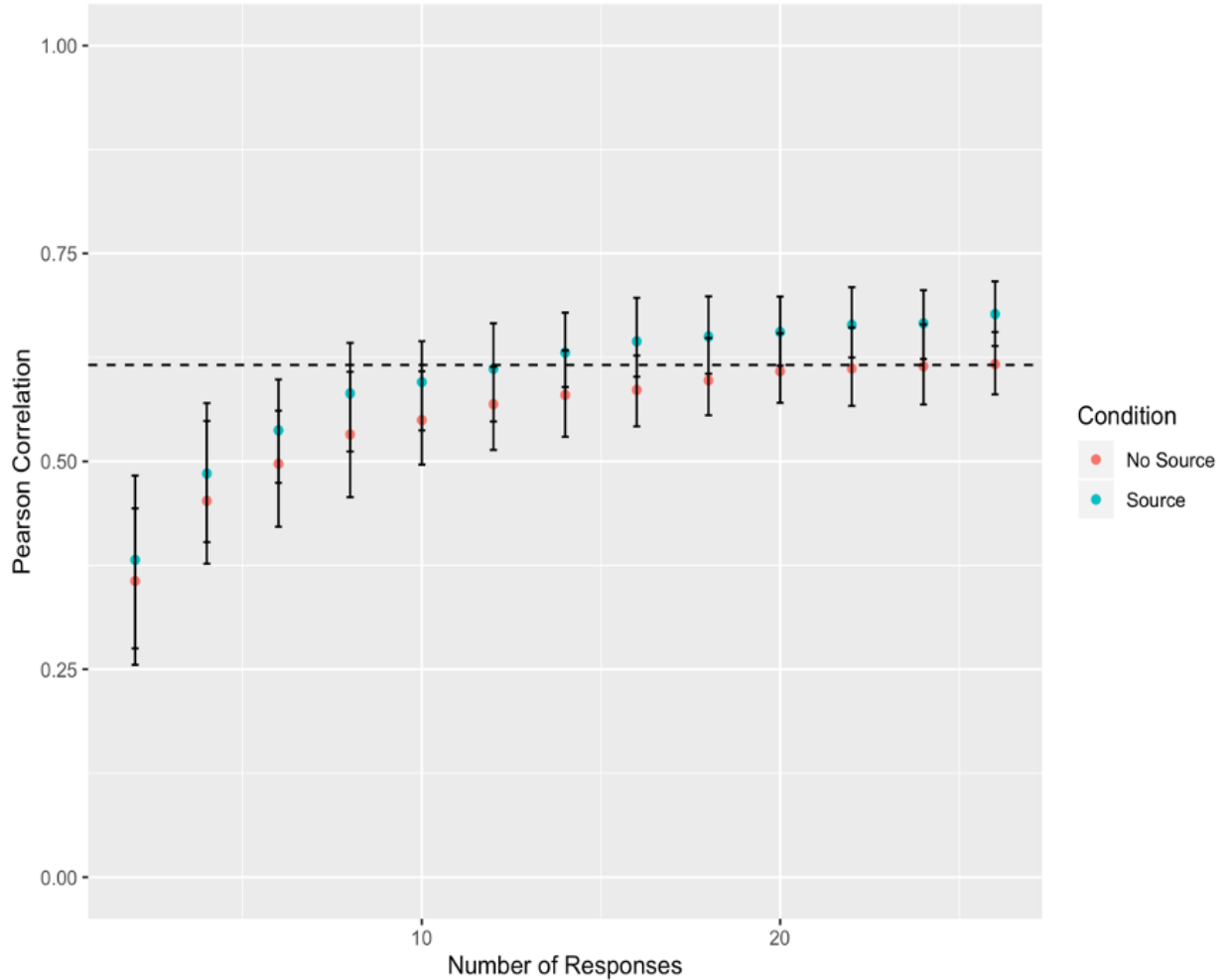


Figure 1.1: Correlation across articles between politically-balanced layperson headline ratings and average fact-checker research-based ratings, as a function of the number of layperson ratings per headline. Laypeople are grouped by condition (Source vs. No Source). The dashed line indicates the average Pearson correlation between fact-checkers ($r = .62$).

1.3.3 AUROC Analysis

While our correlation analysis demonstrates the relationship between fact-checker and crowd ratings and allows for an apples-to-apples comparison of relative performance, understanding how well the crowd’s responses can predict fact-checkers’ binary truth ratings is also a relevant consideration given that most platforms use binary or categorical ratings to flag misinformation content. For this reason, we also use our crowd’s aggregate ratings to predict the fact-checker’s modal categorical rating (which we binarize by giving a headline the label “1” if the modal fact-checker response is “True” and “0” otherwise) and evaluate AUROC of the model. The AUROC can be interpreted in this context as the probability that the crowd will give a higher accuracy rating to a randomly drawn True headline than to a randomly drawn False/Misleading/Couldn’t Be Determined headline.

We repeat the same bootstrap procedure as in our correlation analysis, using a politically

balanced crowd of increasing size n to estimate the AUROC of the given model. The results can be found in Figure 1.2a. As can be seen, the estimate of the AUROC asymptotes with a crowd of around size 26 at .85 for the source condition and .84 for the no source condition ($n = 26$, source condition: AUROC = .85, 95% CI = .83 - .88, no source condition, AUROC = .84, 95% CI = .80 - .86). We can interpret this metric as meaning that 85% of the time, a crowd of size 26 will give a randomly selected “True” headline a higher accuracy rating than a randomly selected False/Misleading/Can’t Tell headline.

Figure 1.2b shows the ROC curves of a crowd of size 26 for the source and no source condition. These curves allow us to evaluate the tradeoff between true and false positives at different score cutoffs. For example, we can see that in the source condition, given a cutoff of 4.5 (slightly above the scale midpoint for the 1 - 7 accuracy rating), we have a false positive rate of 9.5% with a true positive rate of 70%. While ultimately choosing the “ideal” cutoff is a normative and task-dependent question, we believe that these data signal the potential efficacy of using these crowd ratings as features for ranking content in newsfeed or a fact-checker’s queue.

1.3.4 Comparing Crowds of Different Compositions

Finally, we examine how individual differences among laypeople relate to agreement with fact-checker ratings - and whether it is possible to substantially improve the performance of the crowd by altering its composition. In particular, we focus on three individual differences which have been previously associated with truth discernment: partisanship, political knowledge, and cognitive reflection (the tendency to engage in analytic thinking rather than relying on intuition). For each individual difference, we collapse across source and no-source conditions, fix a crowd size of $k = 26$, and examine (i) the correlation between layperson and fact-checker aggregate Likert ratings and (ii) the AUC for predicting whether the fact-checkers’ modal categorical rate is “true” (see Figure 1.3).

As expected, we see clear differences. Democrats are significantly more aligned with fact-checkers than Republicans (correlation, $p < .001$; AUC, $p < .001$); high political knowledge participants are significantly more aligned with fact-checkers than low political knowledge participants (correlation, $p < .001$; AUC, $p < .001$); and high cognitive reflection participants are significantly more aligned with fact-checkers than low cognitive reflection participants (correlation, $p = .01$; AUC, $p = .02$). Strikingly, however, restricting to the better-performing subset for each individual difference does not lead to a significant increase in performance over the baseline crowd on either correlation (Democrats vs. Baseline: $p = .35$, High CRT vs. Baseline: $p = .74$, High PK vs. Baseline: $p = .57$) or AUC (Democrats vs. Baseline: $p = .18$, High CRT vs. Baseline: $p = .59$, High PK vs. Baseline: $p = .60$). While perhaps surprising, this pattern is common to wisdom of the crowds phenomena. The existence of uncorrelated observations from low performers amplifies the high performer signal by canceling out noise. Thus, while it is important that any given crowd includes some number of high performers, it is not necessary to exclude low performers to achieve good performance.

1.4 Discussion

The data we have presented here provide evidence in support of crowdsourcing’s ability to detect misinformation. We find that, given only the headline and lede of an article, a crowd of approximately 10 laypeople can match the performance of fact-checkers researching the full article. We also provide some practical guidance for those wishing to employ such an approach: provide information about the headline’s source. Together, our results suggest that crowdsourcing could be a powerful tool for scaling fact-checking on social media.

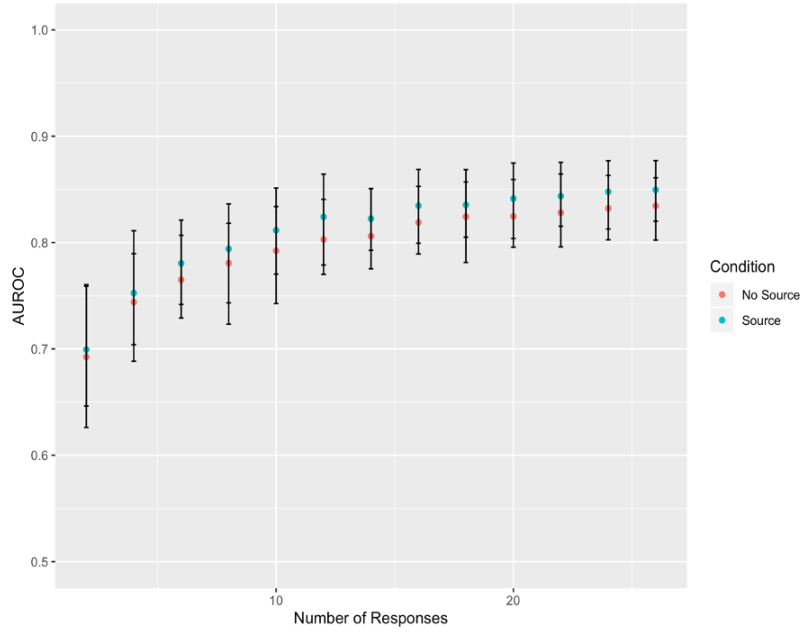
That these positive results were achieved using untrained laypeople without research demonstrates the viability of a fact-checking pipeline that incorporates crowdsourcing. Our results also have practical implications for the manner in which crowdsourcing is implemented. In particular, we advocate for using the continuous crowdsourced accuracy ratings as a feature in newsfeed ranking, proportionally downranking articles according to their scores. A continuous feature incorporates the signal in the crowd’s ratings while guarding against errors that accompany sharp cutoffs of “true” vs. “false”. Additionally, downranking has the benefit of lowering the probability that a user encounters misinformation at all, guarding against the mere exposure effect by which familiar falsities seem true after repetition. While corrections to misinformation have generally shown to be effective [4], [25], the efficacy is dependent on the manner of correction and the possibility of a familiarity “backfire” effect cannot be ruled out [26]. Preventing the spread of misinformation by limiting exposure is a proactive way to fight against fake news.

Despite our positive assessment, we emphasize that our results should not be taken as evidence for replacing current fact-checking efforts. Rather, we see crowdsourcing as just one component of a system that incorporates machine learning, layperson ratings, and expert judgments. While machine learning algorithms are scalable and have been shown to be effective in detecting fake news, they also are domain specific and thus susceptible to failure in a rapidly changing information environment [27]–[31]. Additionally, the limited levels of agreement between our fact-checkers raise concerns about systems that 1) privilege the unilateral decisions of a single fact-checker or 2) use a single fact-checker’s ratings as “ground truth” in supervised machine learning models, as is common practice. Crowdsourced ratings can act as a counterbalance against these other approaches, avoiding the brittleness of software and laboriousness and low fault tolerance of fact-checking.

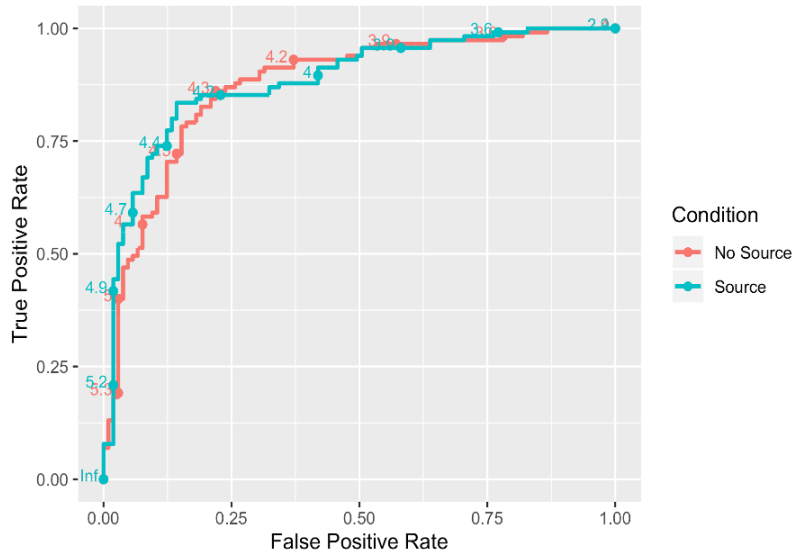
Despite its promise, our study has limitations. First, we note that our results should not be interpreted as evidence that individuals can, in general, accurately identify false information. Even when the crowd performance was good, individual participants often systematically mistook fake news for true and vice versa. Additionally, our results are not generalizable to all situations where misinformation might proliferate. The articles in our stimulus set were published months in advance of the time they were rated by laypeople and fact-checkers. It is possible that under circumstances with rapidly evolving facts, such as in the case of the coronavirus news environment, that results for both the crowd and fact-checkers would differ. Another potential concern is the generalizability of the crowd itself. Mechanical Turk workers differ from the general population in many ways and are more familiar with these types of labeling tasks than other populations might be. Inattentive or, worse, manipulative crowds could potentially negatively affect agreement with fact-checkers.

We suggest recruiting laypeople to rate articles in a distributed manner so as to prevent collusion and preserve the independence of the crowd and including performance checks to mitigate inattentiveness.

In closing, we find promising evidence for the efficacy of using the wisdom of crowds to scale fact-checking on social media. We find that incorporating source information about the article improves the performance of the crowd and discuss the scenarios in which types of crowd provide the best performance. Overall, we believe that in combination with other measures like detection algorithms and trained experts, crowdsourcing can be a valuable asset in combating the spread of misinformation on social media.

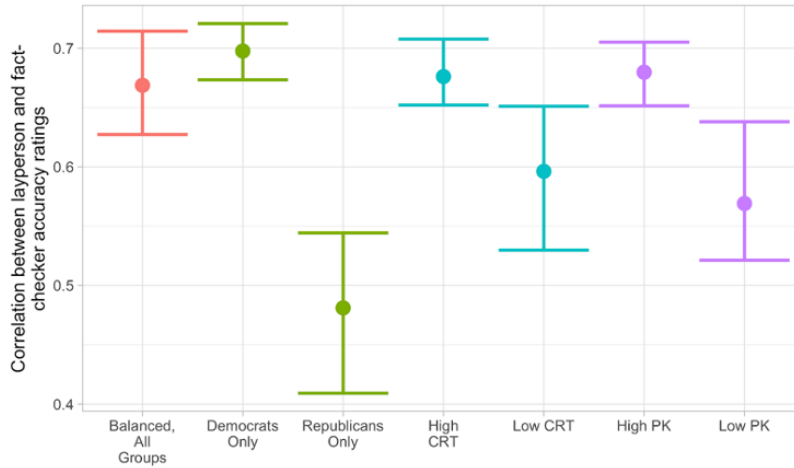


(a) AUROC scores as a function of the number of layperson ratings per headline. AUROC is calculated using a model in which the average layperson headline is used to predict the modal fact-checker categorical rating, where the fact-checker rating is coded as “1” if the modal rating is “True” and “0” otherwise. Laypeople are grouped by condition (Source vs. No Source).

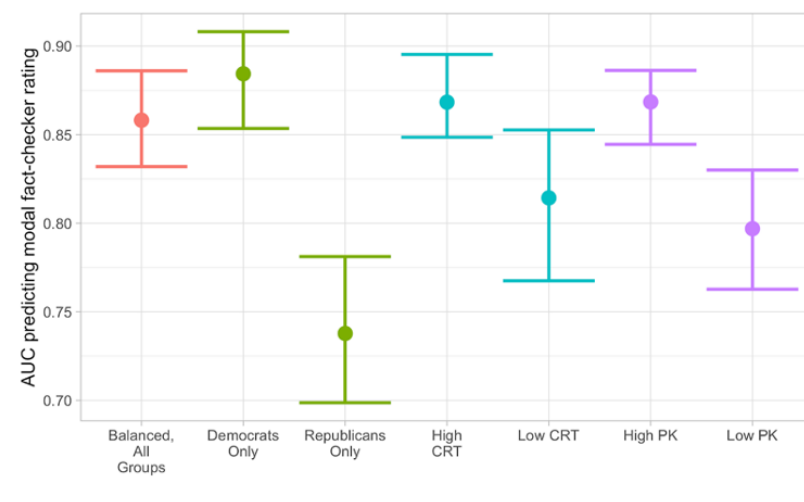


(b) AUROC curves for a politically balanced crowd of 20 laypeople in which the average layperson rating was used to predict the modal fact-checker rating. Laypeople are grouped by condition (Source vs. No Source).

Figure 1.2: Wisdom of Crowds AUROC Analysis



(a) Pearson Correlations between the average rating of a crowd of size 26 and the average fact-checker rating



(b) AUROC for the average rating of a crowd of size 26 predicting the modal fact-checker categorical rating.

Figure 1.3: Comparing crowds with different layperson compositions to a baseline, politically-balanced crowd. For both a) and b), we compare the baseline to a crowd of 1) Only Democrats vs. Only Republicans, 2) Politically-balanced participants with a score above the median on the CRT vs. Those with a score at or below the median CRT, 5) Politically-balanced participants with a score above the median on a Political Knowledge Test vs. Those with a score at or below the median Political Knowledge. Means and confidence intervals are taken from a bootstrap run of 1000 iterations.

Chapter 2

Birds of a Feather Don't Fact-check Each Other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program

1

2.1 Introduction

Understanding the role of partisanship in social media interactions is integral to improving online platforms. For example, partisanship underscores potentially harmful online behavior such as toxic political discourse and harassment of counter-partisan politicians and members of the public [32]–[34]. Exposure to counter-partisan elites via social media can cause increased polarization [35], and more generally, social media use has been causally linked to polarization: being randomly assigned to deactivate Facebook in the leadup to the 2018 U.S. midterm elections significantly decreased polarization [36].

One common explanation for this seemingly toxic political social media ecosystem is the existence of online “echo chambers,” in which users are mostly exposed to content from like-minded others [37]. This idea is largely premised on the observation that people are more likely to be connected to co-partisans online [38], [39], and that shared partisanship causally increases the probability of forming new online connections [40]. Despite being intuitively compelling, however, there is surprisingly little evidence to support the echo chamber hypothesis regarding information exposure. Research finds that connections online are actually less homophilous than offline networks, and the media diets of people on social media are more balanced and moderate than often assumed [41]–[43]. Thus, rather than shielding people from interacting with counter-partisans, there is reason to believe that social media actually increases exposure to counter-partisan content.

As a result, it is of substantial importance for researchers to explore how people react to counter-partisan content when they encounter it online. Studies have shown that people

¹with Cameron Martel and David Rand

are more likely to share news that aligns with their partisanship, regardless of its accuracy [44]–[46], and that politicians’ tweets about members of the other party - which often evoke anger - receive more shares than tweets about members of their own party [47], [48].

These findings are typically interpreted as implying that users judge cross-partisan content negatively. However, it is often extremely difficult to directly assess how social media users actually perceive and evaluate the content they see online. Instead of direct assessments, researchers typically examine on-platform behaviors (e.g. sharing), which are then treated as proxies for agreement. Yet, recent research has shown that there is often a surprisingly large disconnect between sharing and belief [44], [49]–[51]. As a result, the extent to which social media users actually evaluate counter-partisan content more negatively than co-partisan content remains unclear.

In addition to implications for basic research on social interactions and political psychology, understanding whether users judge counter-partisan content more negatively is also important for social media platforms’ efforts to harness the wisdom of user crowds to identify misinformation. Prior work has found that when users are randomly assigned publishers or news headlines to rate, layperson crowds show a high level of agreement with professional fact-checkers [52]–[56] - even when they believe their ratings may influence what content is shown by social media companies [57]. However, if users are free to choose what content to rate, partisanship may lead to systematic biases in what posts are chosen, and what ratings are given. Here, we shed new light on the relationship between shared partisanship and the evaluation of other users’ content. We do so by leveraging data from Birdwatch, Twitter’s recently developed crowdsourced fact-checking platform, which provides clearly quantified data about whether users judge (i) others’ tweets as misleading, and (ii) others’ comments as helpful [58].

2.1.1 The Birdwatch Platform

Birdwatch operates by allowing participants to identify tweets as misleading or not, write free-response fact-checks of tweets, and evaluate the quality of other participants’ fact-checks. When the data for the current research were collected, Birdwatch was in a pilot stage and participation in Birdwatch was available only to a small subset of interested users who applied and were then accepted by Twitter. Twitter aimed to include users from a wide and balanced set of perspectives as pilot participants.

The two main components of Birdwatch are notes and ratings. Notes are the free-response fact-checks participants can write in response to any tweets participants come across and think may or may not be misleading. Notes include various multiple choice questions – most important for this paper is a classification of the tweet as ‘Not misleading’ or ‘Potentially misinformed or misleading’ – as well as an open ended text field where participants can explain their classification and include relevant sources which helped them reach their decision. Participants in the Birdwatch pilot can view notes directly on tweets on their Twitter timeline.

The second main component is ratings, which are evaluations of other Birdwatch participants’ notes. Participants rate the helpfulness of others’ notes, and these ratings are then aggregated by Birdwatch to increase visibility of helpful notes.

For an example tweet, note, and rating aggregation, see Figure 2.1. Birdwatch also

includes a Birdwatch site, where participants and all other Twitter users can view all notes and ratings.

Birdwatch participants can write a note about any tweet they encounter, as well as submit a rating on any note. Additionally, the Birdwatch site has a separate feed of notes which require more ratings for adequate helpfulness aggregation.

2.1.2 The Current Research

In this paper, we examine the relationship between partisanship and behavior on Birdwatch. Importantly, Birdwatch is not focused on political misinformation in particular; Birdwatch users may elect to fact-check any tweet. Furthermore, Birdwatch attempted to reduce partisan motivations by including messaging that emphasized values of building understanding, acting in good faith, and being helpful even to those with whom you disagree. Thus, partisanship may not play a major role in how participants use Birdwatch.

Even if partisanship is associated with fact-checking and helpfulness rating behavior on Birdwatch, it is also unclear a priori what relationships may exist. For instance, users may primarily encounter co-partisan content in their newsfeed, and thus may be more likely to evaluate co-partisan rather than counter-partisan tweets. Alternatively, users may focus on, or even actively seek out, counter-partisan tweets to fact-check. Similar dynamics may also play out for helpfulness ratings: Users may preferentially rate notes by co-partisans as helpful; users could rate counter-partisan notes as unhelpful; or some other combination of evaluations. Thus, examining the partisan dynamics at play on Birdwatch helps inform discussions of partisan behavior on social media, and is critical for understanding how to better implement features such as crowdsourced fact-checking.

To this end, we ask the following research questions:

1. Is shared partisanship an important predictor of whether tweets are rated as misleading, and if so, how?
2. Is shared partisanship an important predictor of whether notes (fact-checks) are rated as helpful, and if so, how?

2.2 Related Work

2.2.1 Partisanship and Online Behavior

A great deal of work has explored the role of partisanship in online behavior. While the concept of online “echo chambers,” which are information environments where consumers are overwhelmingly exposed to confirmatory views [37], has received a great deal of popular attention, academic consensus on the extent to which echo chambers actually exist online is lacking. On the one hand, research has shown that there is substantial ideological clustering on social media sites; that people are more likely to form connections with people who have similar political preferences; and that consumption of political content tends to be more homogeneous than non-political content [38], [40], [43]. Lab studies have also demonstrated

that when given the choice between media outlets, people tend to engage in “selective exposure” and choose to consume content from outlets that align with their political views [59]–[61].

However, observational studies of social media show little evidence for the “echo chamber” hypothesis [62]. Most consumers of news online have relatively moderate political news diets, or otherwise, do not pay attention to news at all [41], [42], [63], [64]. Use of social media has also been found to be associated with increased exposure to counter-attitudinal information, and social recommendations of content online have been shown to blunt the influence of partisan selective exposure [63], [65], [66].

While partisan selective exposure is rarer than expected online, partisan political *behavior* on social media has been robustly documented. Users are more likely to share and retweet content from co-partisans, especially on political topics [38], [43], [67], [68]. Partisans are also much more likely to share fact-checking messages that denigrate their political opponents and boost their political allies [69]. Highly active partisans are also more likely to engage in adversarial interactions with out-party politicians on Twitter [33], [34].

Thus, the majority of empirical work looks at either exposure to, or sharing of, content. Our work contributes to this literature by directly examining *judgments* of content generated by co-partisans versus counter-partisans. People do not necessarily believe much of what they are exposed to [70] or share [44]. Thus, it is an open question as to whether layperson judgments of content will mirror exposure, where we see less empirical evidence of partisan differences than expected, or of sharing, where we see larger effects of partisanship – or show an entirely different pattern.

2.2.2 Motivated Reasoning

There has been a large amount of work in laboratory settings examining partisan judgment of information. The process by which people use biased cognitive processes in order to arrive at a particular directional outcome is called “motivated reasoning,” and many papers have claimed to observe politically motivated reasoning [71]–[73]. For example, an influential study showed that partisans judged confirmatory political claims as higher quality than disconfirmatory claims, and engaged in more counterargument against opposing claims while uncritically accepting supporting claims [74]. This work has also been applied to processing of political misinformation and corrections. Early research showed a “backfire effect,” in which exposure to a correction triggered a counter-argument that actually increased belief in the original misperception [75].

However, recent research has shown that these backfire effects are more likely the exception than the norm, and that corrections typically reduce belief in misinformation on average [76], [77]. Furthermore, studies have shown that even if partisans evaluate co-partisan versus counter-partisan content differently, they might not be exhibiting cognitive bias if partisans have different prior factual beliefs [78], [79]. These different prior beliefs also need not be indicative of less accurate judgments; for example, research has shown that although people are more likely to believe politically concordant news, partisan alignment is not particularly predictive of the extent to which people *differentiate* between false and true news [80], [81].

Importantly, most of these studies use political content that has been hand-picked by experimenters, and thus, may or may not be representative of the content that people actually

encounter online. Therefore, it is unclear to what extent partisanship will play a role in how users evaluate news on social media. Research has found that Twitter is less political than has been typically assumed, and that political content only constitutes a small percent of all tweets – just 13% according to a Pew Analysis [82], [83]. Thus, it is possible that fears of partisan motivated reasoning are overblown and that partisanship will not play a major role in the assessment of content online. By examining the role of partisanship in the judgments of Birdwatch participants, we can shed new light on this question.

2.2.3 Crowdsourced Fact-checking

It is not a priori obvious how these individual-level findings of partisanship translate to group-level assessments of content. Lab studies have shown that small groups of laypeople can generate reasonable levels of agreement with the ratings of experts, including on political content, and that aggregating judgments even among politically homogeneous crowds can lead to more accurate and less polarized judgments [52]–[54], [56], [57], [84], [85].

However, this research was done in settings where laypeople were assigned which pieces of content to rate. In contrast, a study examining editing of Wikipedia articles found that politically homogeneous groups of editors produced worse-quality and less accurate articles than politically heterogeneous groups [86]. On the subject of crowdsourced fact-checking specifically, work in computer science has shown that algorithms where users choose which content to flag could be used to efficiently limit the spread of misinformation, but these algorithms have not been applied in practice or with real user flags [87], [88].

Our work sheds important new light on crowdsourced fact-checking in the wild. Characterizing the behavior of Birdwatch participant crowds who are allowed to choose what to rate illuminates whether partisanship plays a large role in how users 1) rate the accuracy of others’ content and 2) judge the helpfulness of fact-checks.

2.3 Methods

2.3.1 Twitter Datasets

Our analysis of Birdwatch, Twitter’s crowdsourced fact-checking product, used three separate datasets. The first two datasets – the Notes dataset and the Ratings dataset – were provided to us by Twitter, covering all Birdwatch notes and ratings created from the program’s inception on 1/28/21 through 6/29/21. These datasets are very similar to the publicly available datasets found at <https://twitter.com/i/birdwatch/download-data>, except these datasets contained additional information made available for our internal research purposes that allowed us to link the activity of users participating in Birdwatch (the "Birdwatchers") to their Twitter IDs. The third dataset – the Tweets dataset – was collected by us using the Twitter API.

Notes Dataset

The Notes dataset contains the set of 4910 fact-check notes submitted by 1092 unique Birdwatchers. The entry for each note includes the binary classification of the tweet by the

Birdwatcher (either “Not Misleading” or “Potentially Misinformed or Misleading”), the Birdwatcher’s Twitter user ID, the tweet ID of the tweet being fact-checked, and a free-text summary written by the Birdwatcher explaining the rationale for their labeling of the tweet. The Notes dataset was highly imbalanced in terms of classifications: 89.6% of the notes in the sample had a classification of “Potentially Misinformed or Misleading.” Thus, notes functioned largely to flag tweets as potentially misleading. Tweets in this dataset received an average of 1.46 notes (median: 1), and Birdwatchers in this dataset rated an average of 4.5 tweets (median: 2). Full histograms of (a) the number of notes received by each tweet and (b) the number of notes submitted by each Birdwatcher who submitted at least one note can be found in Figure 2.2.

Ratings Dataset

The Ratings dataset contained the 28276 ratings of the helpfulness of the Birdwatch notes, submitted by a set of 2359 Birdwatchers. The entry for each rating includes the note ID, the binary helpfulness rating given to the note (either “Helpful” or “Not Helpful”), and the user ID of the Birdwatcher who gave the rating. The distribution of ratings was much more balanced than the distribution of classifications in the Notes dataset: Of the ratings in our dataset, 65.6% were helpful. Each note received 5.9 ratings from Birdwatchers on average (median: 3), and Birdwatchers rated 12.2 notes on average (median: 4). Full histograms of (a) the number of ratings received by each note and (b) the number of ratings submitted by each Birdwatcher who submitted at least one rating can be found in Figure 2.3.

Tweets Dataset

Finally, we used the Twitter API to pull the full text of tweets about which notes had been written by Birdwatchers, as well as the Twitter user ID of the tweet’s author. Most tweets were accessible via the API; however, some were missing due to the tweet author’s account being suspended or made private, or because the tweet was deleted. At the time of writing, 89.1% of the 3367 total tweets were available for download. Notes for which the original tweets were missing were kept in the dataset, and the relevant tweet-related features were imputed from the means of the data from the existing tweets.

2.3.2 Features

The review helpfulness literature broadly groups features into two different categories: content features, which are features derived directly from the text of the reviews, and context features, which are features like reviewer characteristics that are not derived from the review itself, but nonetheless can be used to predict helpfulness [89], [90]. Drawing on this literature, in our analysis, we determined quantities related to (1) the content of the note summaries and the tweets and (2) contextual features related to the individual characteristics of the tweeters and the Birdwatchers. We then used these features for our main analyses.

Content Features

We extracted several features related to the content of the note summaries and the tweets, which are summarized in Table 2.1. Length, sentiment, and readability have been shown to improve models of review helpfulness in past studies (for a review, see [90]). Additionally, we included the number of URLs as an additional feature, since Twitter suggests that Birdwatch users cite sources in their fact-check notes. We generate the same features for both the note summaries and the tweets.

Table 2.1: Content related features derived from the tweets and Birdwatch notes, respectively.

Feature name	Description
Length	Length (i.e. character count) of the note summary or tweet
Sentiment	Vader Sentiment score from gilbertVaderParsimoniousRule-based2014 for the summary or tweet. [-1,1] scale, where positive values connote positive sentiment.
FK Score	Flesch-Kincaid Reading ease score of summary or tweet. [1,100] scale, with higher values connoting easier reading.
URL Count	Number of URLs in the note summary or tweet.

Context Features

Additionally, we extract several context-related features focusing on user characteristics, which are summarized in Table 2.2. We generate all of these features for the (1) tweeters, (2) Birdwatch note writers, and (3) Birdwatch raters, respectively.

We determine users’ follower count and statuses count (number of posts the user has made) from the the Twitter API. We use the M3Model package [91], a deep-learning model that uses the user’s profile image and textual features of their account, to infer users’ gender and age. Most importantly for our key question of interest, we use the approach of barbera2015tweeting and barbera2015birds, which use the accounts a given user follows to predict their partisanship (Democrat versus Republican), where a score of “0” is represents the partisanship of the median Twitter user. We use this score to assign predicted party identities to users, with scores greater than 0.5 classified as “Republican” and scores less than or equal to 0.5 classified as “Democrat”.

Due to some accounts being deleted, suspended, or made private, we are able to retrieve the full set of user characteristics from 87.7% of tweeters, 92.9% of Birdwatch note writers, and 92.9% of Birdwatch raters. We use mean imputation to fill in any missing data. Descriptive statistics can be found in Table 2.3.

Table 2.2: Context related features derived for the tweet authors, Birdwatch note writers, and Birdwatch note raters, respectively.

Feature name	Description
Follower Count	Number of followers the user has
Statuses Count	The total number of tweets and retweets the user has posted
Age	Predicted age category using M3Model described in [91]. Categories are ≤ 18 , 19-29, 30-39, ≥ 40 .
Gender	Predicted gender using M3Model described in [91]. Coded as "female" vs. "not female."
Partisanship Score	Partisanship inferred using the accounts the user follows, using the method from [43]. [-2.5,2.5] scale, with more positive values indicative of greater affinity for the Republican party.

2.3.3 Models

Our main analyses consist of comparing the performance of various sets of features on two different classification tasks – (1) predicting whether each note classified its respective tweet as potentially misleading and (2) predicting whether each rating rated its respective note as helpful.

We use random forest (RF) models, which have consistently been shown to give good performance on supervised learning tasks that use social media data [92], [93]. In particular, RF models excel at detecting complex interactions between features, which we expect might be relevant when looking at the potential interactions between the partisanship of the tweeter, note writer, and rater. These analyses allow us to measure the maximum predictive ability of a model both in absolute terms and in comparison to the same type model trained on different sets of features, giving us insight into which types of features are most important for our classification tasks.

We performed hyperparameter tuning of the RF model and repeated 5-fold cross-validation (100 times for a total of 500 scores) separately for each of the feature sets. For our evaluation metric, we use the Area-Under-the-Receiver-Operating-Curve (AUC), due to the unbalanced nature of the data and the fact that we value correct prediction of both classes. We report the average AUC and range of the 500 iterations of the cross-validation procedure. Using alternate evaluation metrics like accuracy and F1-score produced substantively similar findings, see Section A.1.1.

Table 2.3: Descriptive statistics for Birdwatch raters, Birdwatch note writers, and tweeters. Gender, age, and party are predicted values derived from machine learning models. Statuses count and follower count are retrieved using the Twitter API.

	Rater	Note Writer	Tweeter
Predicted Gender			
% Female	19%	20%	42%
% Not Female	81%	80%	58%
Predicted Age			
% ≤ 18	22%	17%	4%
% 19-29	30%	30%	15%
% 30-39	24%	26%	24%
% ≥ 40	24%	28%	57%
Predicted Party			
% Democrat	62%	52%	45%
% Republican	38%	48%	55%
Statuses Count			
Mean	17864	25772	57617
Median	6052	8886	16333
SD	38141	53673	111405
Follower Count			
Mean	3584	11069	3959671
Median	386	517	608718
SD	35192	80472	9368102

2.4 Results

2.4.1 Predicting Misleadingness Classification

Random Forest Models

Using the Notes dataset, with tweet and tweeter characteristics merged in from the Tweets dataset, we predict the note’s classification, where “0” corresponds to “Not misleading” and “1” corresponds to “Potentially misinformed or misleading.”

We compare the performance of the RF model predicting note classification using 4 different features sets. First, as a baseline, we train a model using a (1) content-level feature set that contains the features related the tweet’s textual content. Then, we compare the results of this content-only feature set to ones that consist of (2) the partisanship scores of the tweeter and note writer, (3) only the partisanship scores as well as all other (demographic and engagement) context features of the tweeter and note writer, and (4) all features. A description of the features included in each model can be found in Table 2.4.

Table 2.4: Feature sets used to train our model classifying the misleadingness of tweets.

Feature Set	Included Features
Content	Tweet Length, Tweet FK Score, Tweet Sentiment, Tweet URL Count
Partisanship	Tweeter Partisanship Score, Note writer Partisanship Score
Context	Tweeter Partisanship Score, Tweeter Follower Count, Tweeter Statuses Count, Tweeter Age, Tweeter Gender, Note writer Partisanship Score, Note writer Follower Count, Note writer Statuses Count, Note writer Age, Note writer Gender
All	Tweet Length, Tweet FK Score, Tweet Sentiment, Tweet URL Count, Tweeter Partisanship Score, Tweeter Follower Count, Tweeter Statuses Count, Tweeter Age, Tweeter Gender, Note writer Partisanship Score, Note writer Follower Count, Note writer Statuses Count, Note writer Age, Note writer Gender

A comparison on the performance of the RF model predicting whether each note classified its tweet as misleading using the various feature sets can be found in Figure 2.4A. The estimate of the AUC when predicting classification from our baseline content-level feature set is 0.56 (Range = 0.48 to 0.65). This indicates that on their own, the tweet level textual content features we considered do a relatively poor job of predicting which notes classify their tweets as misleading, since the baseline AUC for a model that guesses randomly is 0.5. In contrast, the estimate of the AUC when predicting classifications from just the partisanship scores of the tweeter and note writer is substantially higher, at 0.84 (Range = 0.77 to 0.89). Next, adding additional context features (demographic and engagement features of tweeters and note writers) slightly increased the AUC to 0.87 (Range = 0.80 to 0.92). This suggests

that most of the predictive ability of our context features model comes from tweeter and note writer partisanship scores. Finally, the model using all features has an AUC of 0.85 (Range = 0.79 to 0.91), such that adding content features provided no meaningful benefit beyond context features.

In order to further examine the relative importance of features in our all feature model, we also computed feature importances from one random draw of our cross-validation for the model using all features, which are summarized in Figure 2.4B. In line with our findings from the partisanship features model, we find the greatest feature importance scores for note writer partisanship (0.16) and tweet writer partisanship (0.13). The next most important features were tweeter follower count (0.10) and note writer follower count (0.10). Overall, the results suggest that the partisanship feature set is highly predictive of truth classifications, more so than our baseline tweet content feature set and comparable to a model containing all content and context features.

Logistic Regression Model

One benefit of RF models is that their structure naturally allows them to capture all relevant interactions between features, such as the interaction between the partisanship of the note writer and the tweet writer (i.e. political concordance), allowing them to outperform simple models like logistic regression on classification tasks. However, one drawback of the RF models is that, unlike linear models, it is impossible to identify the direction of the relationship between a feature and the outcome, or to understand which interactions between features are important.

To shed light on these questions, we also conducted a logistic regression model (unregularized) predicting "Potentially misinformed or misleading" classification with standard errors clustered by tweet, tweeter, and note writer, in order to gain insight into the directionality of the important features in our RF models. Our logistic regression model included all features, as well as the interaction between tweeter partisanship score and note writer partisanship score (both z-scored). We include this particular interaction, and not the others, because shared partisanship has been shown to be a relevant predictor of a variety of other social media behaviors (e.g. sharing, following) [40], [43], and exploring whether the same relationship exists in the Birdwatch dataset is a major focus of our paper. Notably, we find a negative interaction between tweeter partisanship score and note writer partisanship score ($b=-1.25$, $SE=0.14$, $z=-9.20$, $p<.001$), such that *shared* partisanship is associated with not giving 'misleading' classifications. Tweeter and note writer follower count were also both negatively associated with 'misleading' classifications ($ps < .026$); for full regression table, see Section A.2.

2.4.2 Helpfulness Classification Results

Random Forest Models

Next, we performed similar analyses predicting whether each rating rated its note as helpful. Using the Ratings dataset, with tweet and tweeter characteristics merged in from the Tweets dataset and note and note writer characteristics merged in from the Notes dataset, we predict

the helpfulness rating of each rating, where “1” corresponds to “Helpful” and “0” corresponds to “Not Helpful”.

Similarly to the truth classification task, we compare the performance of the RF model predicting rating-level helpfulness classification using 4 different features sets. However, the features for this model also include note-level content characteristics and rater-level context characteristics. Thus, we have the following feature sets (1) a content- based features based on textual features of the tweet and note, (2) the partisanship scores of the tweeter, note writer, and rater (3) the partisanship scores as well as other demographic and engagement features of the tweeter, note writer, and rater, and (4) all features. A description of the features included in each model can be found in Table 2.5.

Table 2.5: Feature sets used to train our model classifying the helpfulness of notes.

Feature Set	Included Features
Content	Tweet Length, Tweet FK Score, Tweet Sentiment, Tweet URL Count, Note Length, Note FK Score, Note Sentiment, Note URL Count
Partisanship	Tweeter Partisanship Score, Note writer Partisanship Score, Rater Partisanship Score
Context	Tweeter Partisanship Score, Tweeter Follower Count, Tweeter Statuses Count, Tweeter Age, Tweeter Gender, Note writer Partisanship score, Note writer Follower Count, Note writer Statuses Count, Note writer Age, Note writer Gender, Rater Partisanship Score, Rater Follower Count, Rater Statuses Count, Rater Age, Rater Gender
All	Tweet Length, Tweet FK Score, Tweet Sentiment, Tweet URL Count, Note Length, Note FK Score, Note Sentiment, Note URL Count, Tweeter Partisanship Score, Tweeter Follower Count, Tweeter Statuses Count, Tweeter Age, Tweeter Gender, Note writer Partisanship score, Note writer Follower Count, Note writer Statuses Count, Note writer Age, Note writer Gender, Rater Partisanship Score, Rater Follower Count, Rater Statuses Count, Rater Age, Rater Gender

The findings are summarized in Figure 2.5A. Our baseline content-level model has an AUC estimate of 0.76 (Range = 0.74 to 0.77). This AUC is substantially higher than the corresponding content-only model predicting whether notes classified their tweets as misleading - suggesting that the content features we examine are comparatively more predictive for helpfulness ratings than misleadingness classifications. However, once again our partisanship scores model has a substantially greater AUC estimate of 0.89 (Range = 0.88 to 0.90). And once again, our context features model has an only slightly larger AUC estimate of 0.91 (Range= 0.90 to 0.92). As in our note misleadingness prediction models, these results predicting ratings show that most of the predictive power of the context features model comes

from the partisanship score features. Finally, the all features model has an AUC of 0.92 (Range = 0.91 to 0.93). Thus, although the content features were somewhat predictive on their own, adding them to the context features does not meaningfully improve prediction.

Next, we again examined feature importance from one random draw of our all feature model. As can be seen in Figure 2.5B, the greatest feature importance score is partisanship score of the rater (0.21), followed by partisanship score of the note writer (0.10). Importance scores are also high for number of rater statuses (0.08) and number of rater followers (0.07).

Our helpfulness rating classification model results largely corroborate our main findings from the misleadingness classification models - namely that context features, and specifically partisanship, are highly predictive of both misleadingness and helpfulness ratings.

Logistic Regression Model

We again conducted a follow-up logistic regression model to examine the directionality of the relationships with key features from our RF models. Our logistic regression model included all features from our helpfulness classification all feature model, as well as all interactions between tweeter, note writer, and rater partisanship scores (all z-scored), and clustered standard errors by note, note writer, and rater, in order to predict helpfulness. We find a positive interaction between note writer and rater partisanship score ($b=1.27$, $SE=0.07$, $z=17.02$, $p<.001$), such that shared partisanship between note writer and rater is associated with notes being rated as helpful. We also observe a (somewhat smaller) negative interaction between tweeter partisanship score and rater partisanship score ($b=-0.52$, $SE=0.06$, $z=-8.85$, $p<.001$), such that shared partisanship between tweeter and rater predicts an unhelpful rating; for full regression table, see Section A.2. Given that most note classifications are 'misleading', this pattern of results suggests that raters tend to evaluate notes that agree with their partisanship as helpful, and notes that disagree with their partisanship as unhelpful.

2.4.3 Shared partisanship predicts classifications and ratings

Our results above suggest that shared partisanship is an important feature in both of our models. In particular, the interaction between the partisanship of the tweeter and note writer when predicting the misleadingness classifications, and between the note writer and rater when predicting the helpfulness of ratings are both highly significant and large in magnitude. In this section, we explore those two relationships in further detail.

The relationship between misleadingness classification and the predicted partisanship scores of the note writer and tweeter are shown in Figure 2.6. For clarity, we also summarize the results using the (binary) predicted party of the tweeter and note writer, where values of the political score greater than 0.5 are coded as "Republican," and less than 0.5 are coded as "Democrat", in Table 2.6 [43]. Two findings are important to note. First, Birdwatchers are much more likely to write notes about tweets written by counter-partisans than co-partisans. Predicted Democrats are 3X more likely, and predicted Republicans are 1.5X more likely, to submit a note about a tweet by a counter-partisan than by a co-partisan. Second, while the vast majority of note classifications are misleading, Birdwatchers are more likely to classify a counter-partisan's tweet as misleading than a co-partisan. Predicted Republicans rated 97.2% of tweets by predicted Democrats as misleading (compared to 71.3% by predicted

Table 2.6: (1) Note count and (2) Percent of Notes Rated as “Misleading” for different combinations of the Tweeter and Note Writers’ Predicted Parties (Republican or Democrat)

	Tweeter Democrat		Tweeter Republican	
	Count	Percent Misleading	Count	Percent Misleading
Note Writer Democrat	489	71.3%	1515	95.5%
Note Writer Republican	1003	97.2%	679	82.4%

Table 2.7: (1) Rating count and (2) Percent of ratings that are “Helpful” for different combinations of the Note Writer and Raters’ Predicted Parties (Republican or Democrat)

	Note Writer Democrat		Note Writer Republican	
	Count	Percent Helpful	Count	Percent Helpful
Rater Democrat	9459	83.1%	3017	43.3%
Rater Republican	5609	25.9%	6379	87.1%

Democrats), and predicted Democrats rated 95.5% of tweets by predicted Republicans as misleading (compared to 82.4% by predicted Democrats). Overall, then, Birdwatchers are much more likely to flag counter-partisans’ tweets as potentially misleading.

We see similar evidence of strong co-partisan preference when exploring the relationship between helpfulness ratings and the partisanship scores of the note writer and rater; see Figure 2.7 and Table 2.7. Unlike with note-writing, Birdwatch users – particularly predicted Democrats – rate more notes from *co-partisans* than counter-partisans. Predicted Democrats are 3X more likely, and predicted Republicans are 1.1X more likely, to rate a note from a co-partisan than from a counter-partisan. Second, Birdwatch users are much more likely to classify a co-partisan’s note as helpful than a counter-partisan’s. Predicted Republicans rated 83.1% of notes written by other predicted Republicans as helpful (compared to 43.3% of notes written by predicted Democrats), and predicted Democrats rated 87.1% of notes written by predicted Democrats as helpful (compared to 25.9% of notes written by predicted Republicans).

This preference for concordant notes has implications for the overall average helpfulness ratings of the notes. In Figure 2.8, we show the relationship between the percent of ratings that are from co-partisans and the overall average helpfulness rating of the note, for notes with at least 5 ratings. There is a strong, positive relationship between the percent of co-partisan ratings and the overall helpfulness rating of the note. For a weighted least squares regression of the average helpfulness rating on the percent of co-partisan ratings, where the weights are the number of ratings for that note, the coefficient on percent of co-partisan ratings is .71 ($p < .001$). This means that for every additional 1% percent increase in ratings by co-partisans, the helpfulness rating rises 0.71%. The model has an R^2 of 0.42, meaning that 42% of the variance in helpfulness ratings is explained by the percent of co-partisan raters.

2.5 Discussion

Here we have shown that shared partisanship is an important predictor of how Twitter users in the Birdwatch program evaluate others' posts, with tweets from counter-partisans judged as more misleading than tweets from co-partisans, and notes (e.g. fact-checks) from counter-partisans judged as less helpful than notes from co-partisans. We add to the literature on partisan selective exposure and partisan selective sharing by demonstrating a related phenomenon: partisan selective *evaluation*. These findings are notable and perhaps surprising, since much of the content on Twitter is not political in nature, and political content is a fairly small subset of most viral forms of misinformation on social media [82], [94]. It was not a priori obvious, therefore, that partisanship should be such a predictive factor when judging the accuracy of tweets or the helpfulness of fact-check notes – especially when compared to other theoretically relevant features, like the number of sources cited in the note.

Given these findings, it is possible that partisanship is motivating users to volunteer for, and contribute to, Birdwatch in the first place. Other research on crowdsourcing for citizen science has found that extrinsic motivations (e.g. status markers within the community) and intrinsic motivations (e.g. belief in the overall goal of the project and individual level interest) are important motivations for participating in these types of project [95]–[97]. Partisanship could play into both types of motivations in the case of Birdwatch. In terms of extrinsic motivations, it is possible that the helpfulness feedback system is signalling to Birdwatchers that partisan-aligned political content is most valued by other Birdwatchers, since the pattern of helpfulness votes suggests a partisan cheerleading effect. As for intrinsic motivations, research has shown that partisanship is a highly salient and important part of people's identities [98]. It is therefore possible that people are participating in Birdwatch to either advance their partisan views, regardless of truth; or because they are genuinely concerned about misinformation generated by users from across the political aisle. Even though viral misinformation is to a large extent non-political in general, it is possible that Birdwatch users are particularly motivated to fact-check partisan information because partisanship is an important part of their identities. Past work has shown that evaluations online are costly to provide and thus scarcer than optimal, so partisan motivations might actually be beneficial for soliciting notes and ratings in a non-paid platform like Birdwatch [99]. Indeed, past research on Wikipedia suggests that editors who are more politically extreme are more willing to spend time and effort advocating for their viewpoint on Wikipedia articles, and thus, some level of "bias" among editors might spur an optimal level of debate and activity on the platform [86]. A similar dynamic could be happening on Birdwatch.

Importantly, the preferential flagging of counter-partisan tweets we observe does not necessarily impair Birdwatch's ability to identify misleading content. It is possible that partisans are successfully identifying misinformation from across the aisle (even if they are not scrutinizing content from their own co-partisans as closely), and/or that aggregating ratings from the entire community cancels out bias from both sides (as in [57]). Consistent with this possibility, a preliminary investigation found that among 57 tweets which a majority of Birdwatchers flagged as misleading, 86.0% were also rated as misleading by at least one of two professional fact-checkers (recruited from [52]). Future work should investigate these issues more thoroughly by assessing the veracity of the full set of fact-checked tweets relative

to some ground truth (e.g. by having professional fact-checkers evaluate all tweets).

Beyond the specific use case of crowdsourced fact-checking on Twitter, our study contributes to research on partisanship and misinformation more generally. Our observation that Birdwatch participants were much more likely to choose to fact-check counter-partisan tweets provides ecologically-valid support for previous findings from survey experiments suggesting that people subject out-partisan content to more scrutiny than in-partisan content. For example, Taber and Lodge [74] found that partisans scrutinized counter-attitudinal content far more closely than pro-attitudinal content, which they did not critically examine. In their work, exposure to opposing arguments led to ideological polarization rather than moderation, and although we do not measure polarization as an outcome, it is possible that a similar phenomenon could happen in this instance.

Interestingly, the pattern of partisan selection evaluation that we observe on Birdwatch cannot be explained by partisan selective exposure. We inferred user’s partisanship based on the accounts they followed, and thus, by construction, users feeds were more likely to contain co-partisan content than counter-partisan content. Nonetheless, users were more likely to post fact-checks of counter-partisan tweets, and, conditional on performing a fact-check, more likely to rate counter-partisan tweets as misleading. Furthermore, such partisan selection is likely driven primarily by disagreement with and motivation to fact-check (potentially misleading) counter-partisan content itself, rather than motivation to fact-check based on partisan account cues. This is because the partisanship of profiles is likely opaque to users, with some notable exceptions such as accounts of politicians and other political elites. Thus, it is likely that counter-partisan cues in tweeted content itself is motivating partisan fact-checking.

While Birdwatch notes were only viewable by the public on a separate website at the time these data were generated, Birdwatch pilot users could see helpful notes attached to the tweets in their feeds. With this in mind, it is important to consider that the users who followed accounts with a similar political lean to a given tweet’s author – and who presumably are thus more likely to come across the tweet organically in their newsfeed – were more likely to be critical of (i.e. rate as unhelpful) notes that marked the tweet as misleading. Thus, the most likely potential consumers of the fact-check were least likely to consider the fact-check helpful. This negative assessment could have important implications for polarization, especially if the fact-checks in question bear more resemblance to partisan “dunking” than to corrections by fact-checkers [47], [100]. While research has shown that fact-checks – even partisan ones – generally decrease belief in misinformation [76], [77], other work has shown that both exposure to counter-partisan content and negative characterizations of counter-partisans can increase polarization [35], [100], [101]. Public corrections could also cause backlash from the original tweeter, as has been shown in a field experiment on Twitter where replying to a misinformation tweet with a link to a fact-check increased the partisan slant and toxicity of the original tweeter’s subsequent retweets [102].

Furthermore, the partisan behavior we observe has important implications for the ability of the Birdwatch helpfulness rating system to identify helpful fact-checks. Both our paper and work by others [103] has identified substantial partisan herding in Birdwatch ratings, identifying a potentially substantial flaw in the helpfulness rating system. Perhaps due to these problems, Twitter has been implementing changes to the rating system. While the data analyzed here were being collected, Twitter labeled notes that had at least five ratings and



an average helpfulness score of at least 0.84 as “Currently rated as helpful” and highlighted these notes more prominently on their site. Subsequently, in June 2021, Twitter changed their helpfulness labeling algorithm to weight notes by a Birdwatcher’s reputation, which is derived in part based on the agreement with consensus rating of the notes they rated in the past [104]. However, if, as we see, Democrats are less likely to submit notes for counter-partisan content than Republicans, then Democrat raters could have a higher reputation due to their greater willingness to engage in partisan cheerleading – rather than higher overall quality. On the other hand, if Twitter just does a simple aggregation of helpfulness scores without reputation rating, they risk a situation where partisan herding could lead to actually helpful notes getting downvoted and unhelpful notes getting boosted due to brigading. It is possible that a different aggregation methodology for helpfulness that balances ratings from parties could prove beneficial, or that Birdwatch should dispense with the helpfulness ratings entirely and instead only focus on classifying tweets as misleading, and/or providing fact-checking notes.

There are important limitations to our study. We cannot identify from these data whether the pattern we observe is the result of politically motivated or otherwise biased reasoning. The observed pattern could also be explained by partisan difference in prior factual beliefs, leading (rational Bayesian) partisans to be more likely to fact-check out-party content simply because it is surprising, rather than because of a political vendetta or bias [105]. Furthermore, it is important to note that the Twitter users who opted in, and were subsequently selected, to participate in the Birdwatch pilot are surely unrepresentative of Twitter users in general, or of Americans more broadly. Men outnumber women 4:1, and the average tweet count Birdwatchers (>25,000) suggests that the users were quite active on Twitter. Additionally, they may be more politically engaged and extreme - and thus more responsive to shared partisanship - than the average Twitter user. Future work should examine how Birdwatchers compare to more representative populations, and evaluate what individual differences predict the relationship between shared partisanship and choosing to rate others’ content. Future research should also examine the extent of partisan herding in Twitter replies more generally, rather than just on Birdwatch. Such analyses may shed light on how similar the partisan dynamics observed in a crowdsourced fact-checking setting are to partisan dynamics on Twitter overall. It will be informative to see whether partisan herding is exacerbated by a fact-checking directive, or if similar partisan communication patterns exist (perhaps to an even greater degree) on Twitter outside of Birdwatch.


Furthermore, we recognize that one potential drawback to this research is that, for privacy reasons, we cannot release IDs of the Twitter accounts participating in Birdwatch that were used in our analysis. For transparency, we have posted our code on OSF: <https://osf.io/acx3j>. Twitter has been releasing anonymized datasets of the notes and ratings from Birdwatch on their site <http://twitter.com/i/birdwatch/download-data> and if de-identified datasets including the relevant co-variables from our analyses become available we will add them to OSF.

In sum, we have shown that shared partisanship is a strong predictor of whether a user rates a tweet as misleading or a fact-check as helpful in the context of Twitter’s crowdsourced fact-checking platform Birdwatch. While we do not believe that our findings mean that social media platforms should abandon crowdsourcing as a tool for identifying misinformation, the patterns we observe clearly indicate that it is essential to consider partisan dynamics when


designing crowdsourcing systems.

A  **Amazon News**  @amazonnews · Mar 24
 Replying to @repmarkpocan
 1/2 You don't really believe the peeing in bottles thing, do you? If that were true, nobody would work for us. The truth is that we have over a million incredible employees around the world who are proud of what they do, and have great wages and health care from day one.

14.3K 19.4K 4.7K

B  **Currently rated helpful** ...
 Informative · Cites high-quality sources

Potentially misleading Mar 25
 Amazon has a documented history of labor violations, including pushing employees to work so much they do not have time to use the restroom.
<https://www.theguardian.com/technology/2020/feb/05/amazon-workers-protest-unsafe-grueling-conditions-warehouse>
<https://www.newsweek.com/amazon-drivers-warehouse-conditions-workers-complains-jeff-bezos-bernie-1118849>
<https://www.npr.org/2020/07/31/897836765/amazon-workers-respond-to-jeff-bezos-testimony-before-congress>

C  **Currently not rated helpful** ...
 Sources not included or unreliable · Misses key points

Potentially misleading Mar 25
 Amazon workers are treated awfully. Thank you.

Figure 2.1: An example of two Birdwatch notes, along with the focal tweet. (A) is the tweet that has been flagged by Birdwatch users. (B) is a Birdwatch note which labels the tweet in (A) as “Potentially Misleading”. The note shown in (B) has been labeled as “Currently rated helpful” by Twitter, based on the high aggregate helpfulness rating given to it by other Birdwatch users. (C) is a Birdwatch note that also labels the tweet in (A) as “Potentially Misleading”. However, the note in (C) has been labeled as “Currently not rated helpful” by Twitter, likely based on the low aggregate helpfulness rating given to it by other Birdwatch users.

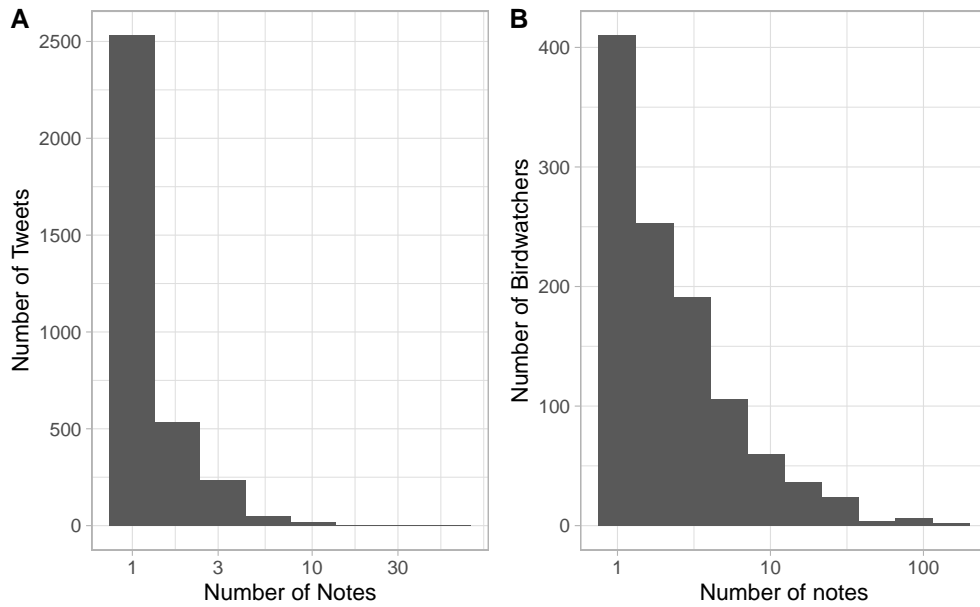


Figure 2.2: (A) A histogram of the the number of Birdwatch notes received by each tweet. The histogram is long-tailed, and most tweets receive one note, although some receive up to 30. (B) A histogram of the number of notes submitted by each Birdwatcher who wrote at least one note. The histogram is also long-tailed, with most Birdwatchers submitting less than 5 notes, but some submitting more than 100. Note that for both histograms the X-axis is on a logarithmic scale.

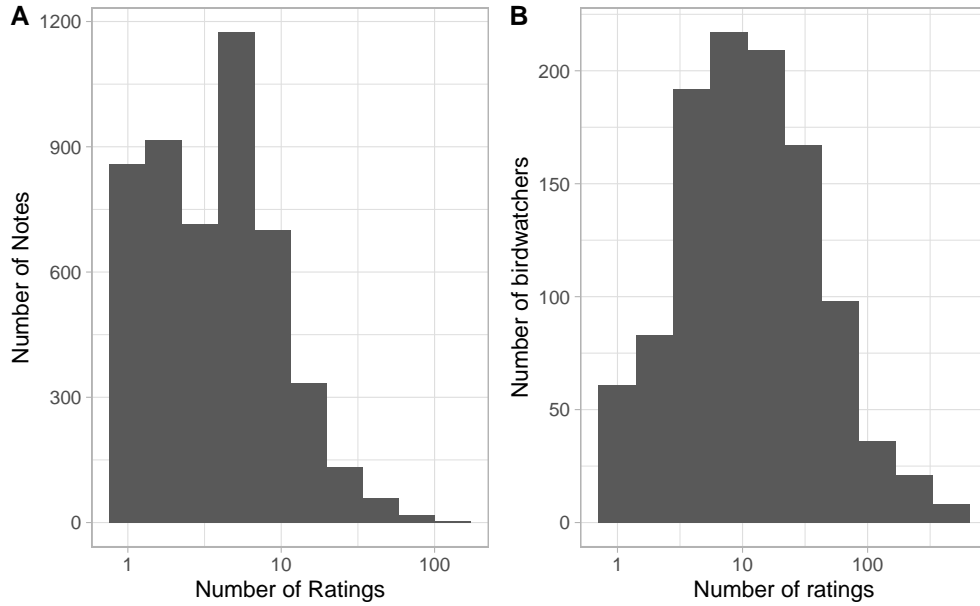


Figure 2.3: (A) A histogram of the the number of ratings received by each Birdwatch note. Most notes receive 10 ratings or less, with some notes receiving up to about 100.(B) A histogram of the number of ratings submitted by each Birdwatch user who submitted at least one rating. Note that for both histograms the X-axis is on a logarithmic scale.

Histograms of the number of ratings submitted for each note and the number of ratings submitted by each Birdwatch user

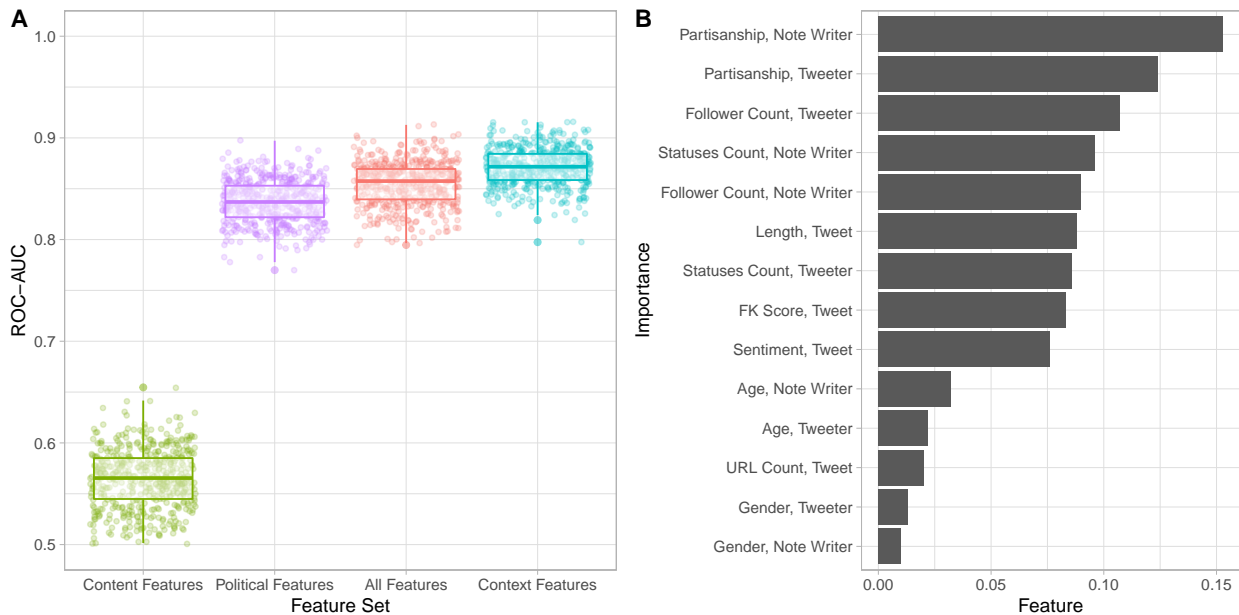


Figure 2.4: (A) Comparing the performance of RF models predicting the misleadingness classification of tweets with different feature sets (B) Comparing the relative importance of features for an RF model trained with all features

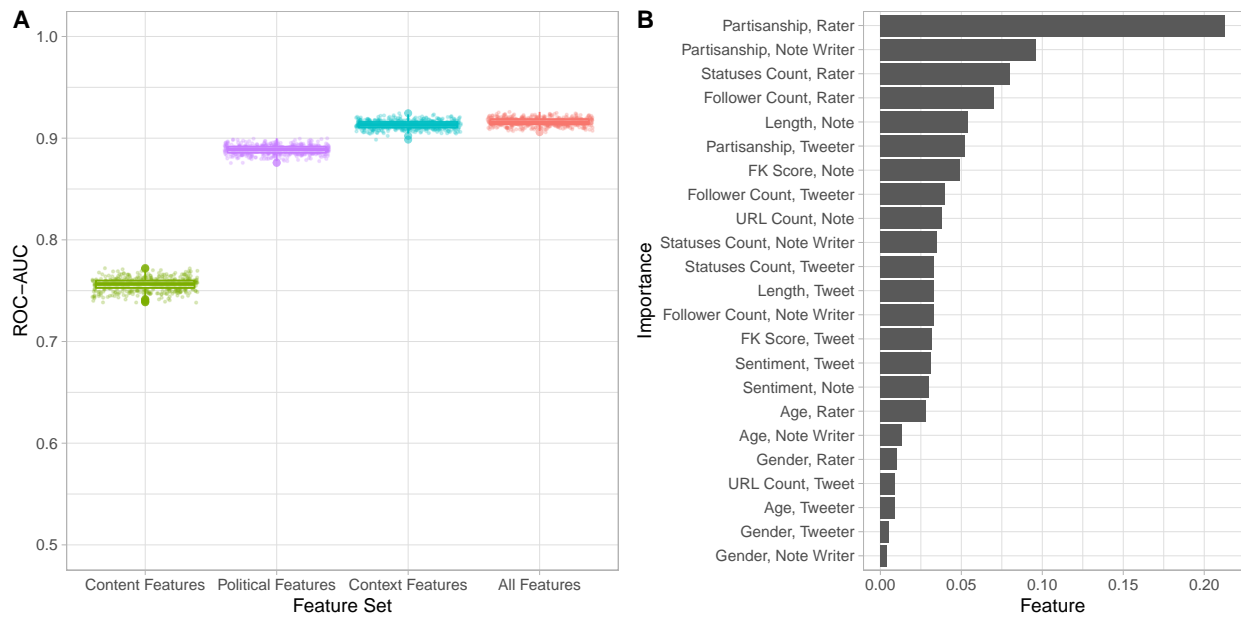


Figure 2.5: (A) Comparing the performance of RF models predicting the helpfulness ratings of Birdwatch notes with different feature sets (B) Comparing the relative importance of features for an RF model trained with all features

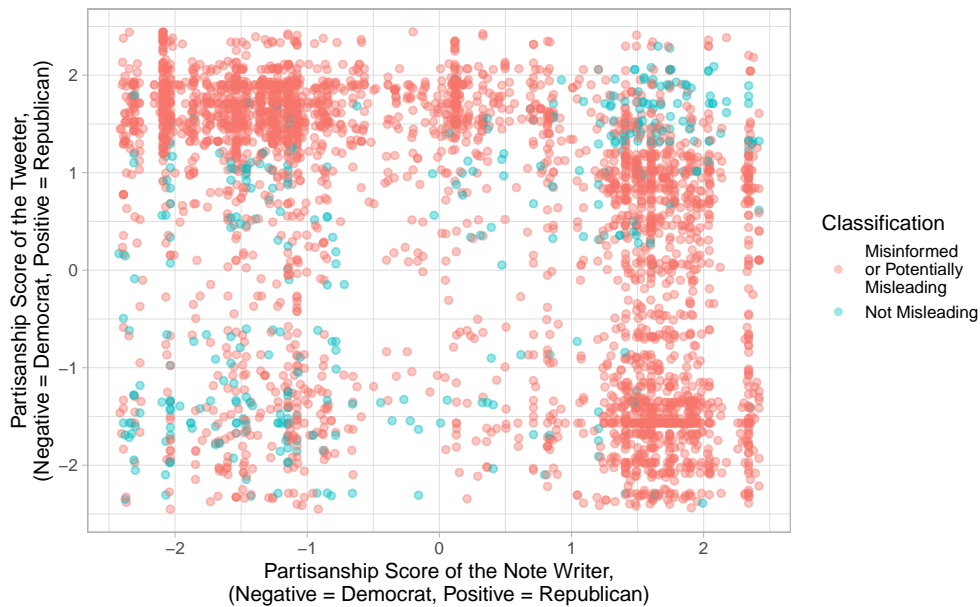


Figure 2.6: Misleadingness classifications of tweets, by the partisanship score of the note submitter and the partisanship score of the tweeter. Each point represents one tweet.

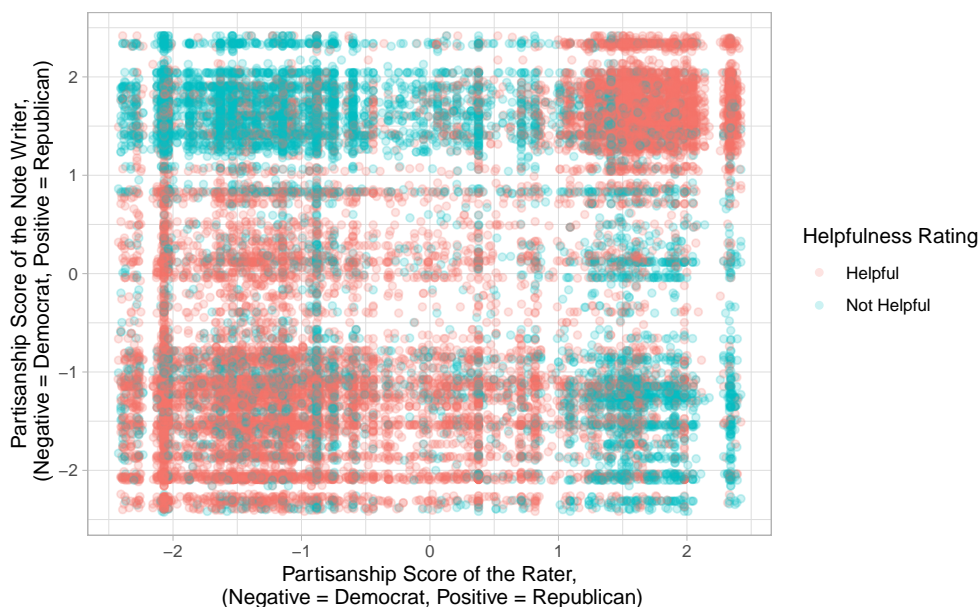


Figure 2.7: Helpfulness ratings of notes, by the partisanship score of the rater and the note submitter. Each point represents one note.

Scatter plot of note helpfulness ratings with partisanship score of the rater on the X-axis and partisanship of note submitter on the y-axis

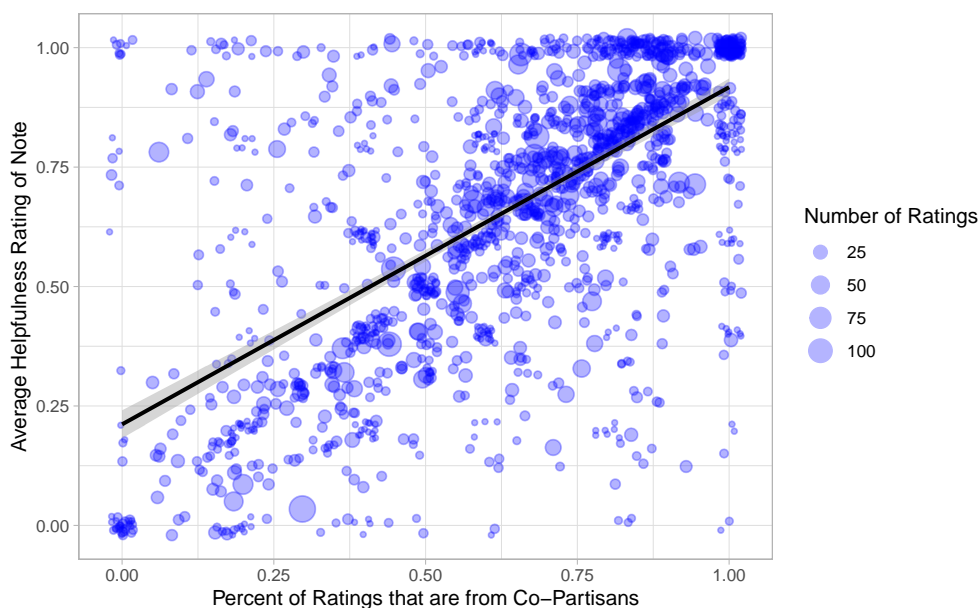


Figure 2.8: Predicting helpfulness ratings as a function of the percent of ratings that are from co-partisans for notes with greater than or equal to 5 ratings. Each point represents one note, where points are sized by the number of ratings received by each note.

Chapter 3

Quantifying the Impact of Misinformation and Vaccine-Skeptical Content on Facebook

3.1 Introduction

In recent years, the spread of misinformation online has become a key concern for policy-makers and the public, and a major focus of study for researchers [106], [107]. This attention is largely motivated by the association between misinformation and real world impact, including phenomena like the January 6th Capitol Hill Riots and the rejection of public health messages during COVID-19 pandemic. Yet, despite a wealth of research on the viral spread of misinformation [88], [91], [108]–[110], the prevalence of fake news [10]–[12], [111] and the cognitive psychology driving sharing of and belief in falsehoods [80], [112]–[114], consideration of the real-world causal impact of misinformation has been largely relegated to introductory paragraphs and discussion sections [115]. Little work has demonstrated a plausible causal pathway between exposure to online misinformation and large-scale social harm.

Under what circumstances could online misinformation have had the sweeping societal impact put forth by researchers and media critics? We posit that for any type of information to have widespread impact on people’s behavior, it must meet two criteria. First, people must see it. Second, it must influence behavior, conditional on being seen. In short, we define impact as the product of exposure and persuasive influence. That is, harmful misinformation which no one sees is not impactful; nor is misinformation that is widely seen but irrelevant to people’s decision-making. While a great deal of past research has focused on prevalence by examining social media trace data, only a handful of studies have examined the effect of misinformation exposure using lab experiments (with conflicting results, see [116] for a review) – and to our knowledge, no studies have yet tried to meaningfully link these two sides of the equation. Thus, whether and to what extent misinformation has actually had an important impact on society remains an open question.

Our work addresses this gap. We propose a framework for quantifying the impact that misinformation (or any other form of persuasive content) has on societal outcomes by combining exposure data with causal estimates of persuasive influence. We apply this novel

methodology towards a question of broad social interest: assessing the impact that content on Facebook had on people’s intentions to take a vaccine for COVID-19. By combining i) data about the exposure to vaccine URLs on Facebook with ii) results from experiments measuring the effect of different vaccine-related headlines on vaccination intentions, we estimate the overall impact of Facebook on vaccination rates in the US. We then compare the impact of misinformation with other types of vaccine content to assess whether false content had a disproportionate negative effect on COVID-19 vaccine uptake in the US.

We choose to focus on vaccine content because intentions to take a COVID-19 vaccine is a measurable outcome that has been shown to correspond to the real-world harm of vaccine refusal [117]. Since the COVID-19 vaccine has been shown to be safe and effective in preventing the contraction and spread of COVID-19, lowering the rate of vaccination directly impacts social welfare through increased sickness and death [118], [119]. Fewer than 70% of Americans have been fully vaccinated for COVID-19, a rate lower than similar developed countries [120]. Therefore, understanding the reasons for this vaccine refusal has implications for this and future public health emergencies.

We consider the role of vaccine misinformation in particular because even though the “infodemic” of viral falsehoods is frequently cited as an obstacle to the adoption of public health measures, little work has been done to show a causal connection. President Joe Biden claimed Facebook was “killing people” by letting anti-vaccine misinformation spread on the platform, and FDA commissioner Robert Cardiff said he “believe[d] that misinformation is now our leading cause of death” [121]. However, despite a wealth of academic research studying the correlational relationship between social media use, endorsement of vaccine conspiracies, and vaccine hesitancy [122]–[125], there has been little “gold-standard” experimental work testing the impact of misinformation on vaccination intentions [126]. The few lab studies testing for a causal relationship between vaccine misinformation and behavioral intentions have shown conflicting evidence [127], [128], and research has found that the most already vaccine-resistant people selectively seek out the most misinformation content online [129]. Past work has also posited that content that is “vaccine-skeptical”, defined as content that “could undermine faith in approved vaccines even if [it does] not reflect an explicitly anti-vaccine viewpoint,” could play an important role in driving vaccine hesitancy even if it is not outright false or intentionally trying to increase vaccine refusal [129].

Despite these open questions of causality, significant resources have gone into finding solutions for the misinformation problem. These include financial resources, e.g. tens of millions of dollars of grants from the CDC and social media companies, as well as research efforts [3]. Academics and practitioners have tested a wealth of potential interventions to fight COVID misinformation on social media, including automated tracking of falsehoods, prebunking and debunking, and accuracy nudges [30], [51], [125], [130]–[132]. However, one barrier to evaluating such interventions is that even though the motivation of these interventions is to decrease vaccine refusal, most research instead measures the spread of or belief in misinformation as the outcome of interest. In light of the complicated relationship between exposure, belief, and action, we argue these measures are likely misspecified when considering questions of behavior. Lowering the number of fake URLs on Facebook is ineffective if people do not see them or change their minds about vaccination because of them.

Here, we take a different approach. Rather than using sharing or belief in misinformation as a proxy for harm, we directly measure harm as decreased willingness to take a COVID

vaccine. While we treat misinformation with particular scrutiny, we also model the impact of all vaccine-related link content popular on Facebook on vaccine refusal. By taking an a priori agnostic view of what content might change vaccination intentions, we discover from the bottom-up which types of content drive overall vaccine hesitancy, and then quantify how much of this vaccine-skeptical content is outright misinformation.

Our paper proceeds as follows. First, we consider which types of vaccine content changed willingness to take a COVID-19 vaccine, conditional on exposure, using two online survey experiments. One explanation for conflicting evidence on the causal impact of misinformation on vaccination intentions is that the researchers used different stimuli to operationalize the shared concept of “misinformation”—sometimes called the “stimulus-as-fixed-effect-fallacy” [133], [134]. Researchers treat misinformation as a uniform category of content, even though false claims can vary wildly. For example, “COVID-19 is only as deadly as the seasonal flu” vs. “A pod of humpback whales returned to the Arabian Sea offshore from Mumbai, India following the COVID-19 lockdown” were both claims labeled as “false” by experts, but with very different implications for public health [135]. Furthermore, factually-accurate content might also increase vaccine hesitancy, e.g. news of the governments pausing rollout of the JJ vaccine following blood clot risks [136]. To broadly understand what content drives vaccine hesitancy, we tested the effect of 130 different vaccine treatments—both misinformation and factually-accurate information—on intentions to take a COVID-19 vaccine. Then, by examining the heterogeneity between different stimuli, we identify which content features beyond just accuracy can lower willingness to take a vaccine. In particular, we found that content that suggests the vaccine was harmful to a person’s health was particularly damaging.

Next, we consider questions of exposure. Again, we find conflicting accounts of the prevalence of vaccine misinformation on social media. Some research supports the “infodemic” framing, identifying cases where viral COVID vaccine misinformation shared by a small number of misinformation “superspreaders” generated millions of interactions on social media [137]–[140]. However, other work has shown that fake news sharing and consumption is comparatively infrequent and highly concentrated among the heaviest news consumers [11], [12], [52], even in the context of COVID-19 [111]. Yet, none of this prior work has been able to observe the actual views received by specific content on social media, instead relying on proxies such as the number of shares or traffic to a certain domain. In contrast, our work uses a large-scale dataset released by Facebook Social Science One dataset to measure the actual views received by individual URLs on Facebook [141]. For the first time, we can calculate the number of times that specific vaccine-related URLs were seen by US Facebook users, and compare exposure to false content with exposure to other types of vaccine-related content.

In the third section, we extrapolate the results from our survey experiments to the larger Facebook URL dataset. Crowdsourcing has been shown to be a powerful tool in both identifying misinformation on social media [52], [54], [87] as well as judging the persuasive effects of nudge messages [142]. It has also shown promise in improving robustness of machine-learning models across time, since humans can apply prior knowledge that can correct for changes to data that are not observable to models [143]. Here, we show that crowdsourcing, in combination with transformer-based natural-language-processing models, can be used to predict which content is most likely to decrease willingness to take a vaccine. As a result, rather than relying on a URL’s source quality or fact-check status as a surrogate measure

for harm, we directly estimate the harm caused by each individual URL.

We conclude by estimating the overall impact of Facebook content on vaccination intentions by taking the product of exposure and our predicted persuasive effect for each URL. This process allows us to assess the impact of vaccine misinformation, compared to other vaccine content, on vaccination intentions during the initial rollout in the first quarter of 2021. To preview our results, we find plausible evidence that content on Facebook had a significant negative impact on vaccination intentions in the US. However, due to limited exposure, outright false misinformation made up only a small fraction of the total impact. In contrast, vaccine-skeptical mainstream media articles highlighting rare cases of deaths following COVID vaccination had far greater exposure - and thus were much more negatively impactful. We consider the ramifications of these findings for social media platforms, journalists, and researchers.

3.2 Methods

3.2.1 Facebook Exposure Data

Facebook and Social Science One URL Shares Dataset

We used Social Science One and Facebook’s URL Shares Dataset to identify URLs related to the COVID-19 vaccine during the initial vaccine rollout during the first 3 months of 2021[141]. This dataset contains information on all URLs publicly shared at least 100 times on Facebook and covers data from 2017-2022, at the time of writing. The dataset includes

1. Descriptive characteristics of URLs, like the title, description, domain, and third-party fact-checker ratings of the URLs and
2. Counts for actions taken on each URL including views, clicks, and shares for each URL, grouped by URL-demographic-month bucket.

Facebook de-duplicated each action such that the engagement metrics reported are not the total number of views (or shares, etc.), but rather the total number of users who viewed (or shared) the URL. For example, a given row might describe the number of times a particular URL was clicked, shared, and viewed by women in the U.S., aged 18–24, who lean conservative, in January 2022.

Facebook also added differentially private noise to the engagement-related columns of the dataset [144]. While this noise can change the results of many statistical procedures, the sums of differentially private columns are unbiased estimates of the true sums and thus, we did not do any further corrections in our analysis. However, we do calculate the confidence intervals for each top story, which are proportionally very small (see Guess, Aslett, Tucker, *et al.* [145] for a reference for the calculation of these confidence intervals and deeper discussion of the URL Shares dataset).

Research has shown the 100 public-share threshold can lead to biased conclusions when comparing shares of high-engagement vs. low-engagement content (e.g. the share of clicks to fake news domains vs. non-news domains)[146]. However, since our analysis is largely

concerned only with the top-viewed stories, this bias is unlikely to change our results. If anything, the threshold would bias the results in favor flagged misinformation – which is known to generate more engagement relative to exposure than non-misinformation content (see [146]).

Our universe of 13,206 vaccine-related URLs were gathered using the following procedure. We queried the full URL dataset for all URLs that were i) posted for the first time between 1/1/21 and 3/31/21, ii) had “vaccin*” or “vax” in their title, description, or URL, and iii) were primarily popular in the US. To identify URLs primarily popular in the US, we subsetted to URLs that had greatest number of public shares in the US (as opposed to any other country) using the “public_shares_top_country” field in the URL Shares dataset.

We excluded the 26 URLs that had missing headlines and descriptions. For each URL, we calculated the number of views in the US from the month it was first posted and one additional month. We use this sliding window to reduce the proportion amount of differentially-private noise for each URL, since the amount of noise is constant with each month of a URL appearing in the dataset, while the number of URL actions (likes, shares, views) the URL garners tapers off very quickly with time.

This dataset also includes information about whether the URL had been fact-checked by third-party fact-checkers. URLs sent to fact-checkers could be labeled as either ‘True’, ‘False’, ‘Mixture’, ‘Missing Context’, or ‘Not Rated’. URLs labeled as ‘Not rated’ were sent to fact-checkers but were not subsequently rated; URLs that were not sent to fact-checkers at all had a rating of ‘NA’. According to Facebook, content rated as ‘False’ or ‘Mixture’ – but not ‘Missing Context’ are demoted in feed. More information on Facebook’s third-party fact-checker can be found in the URL shares dataset documentation [141].

According to Facebook, the algorithm that flags content for fact-checking is based on signals such as the number of times a URL has been shared or whether it contains keywords associated with false stories, among other signals. Fact-checkers are encouraged to prioritize content that is flagrantly false [1].

Facebook Headline Clustering

Because many of the headlines are AP reprints and variations of the same news event, we cluster the headlines into “stories” for better interpretability. Aggregating stories together also helps reduce differentially-private noise without sacrificing high level insights about which stories were most popular during this time. We implement the following clustering procedure. First, we stemmed the words using a Snowball Stemmer and removed English stopwords and punctuation. Then, we used k-means clustering with $k=500$ on the tf-idf scores. We chose a high number for K to maintain a relatively high level of purity within clusters, such that only the most similar headlines referring to the same events (e.g. “Florida doctor dies after taking COVID-19 vaccine”) or close variations on stories (e.g. “Severe side effects of the second dose”) are clustered together. We exclude the largest cluster, containing 516 headlines, since inspection revealed that these URLs were largely unrelated. We choose the headline nearest to the center of the cluster to be the “representative” cluster headline.

We also give examples of the top URLs without clustering. This methodology does not change the interpretation of results. These robustness checks can be found in Section B.0.8.

Low Credibility Domains (Exposure)

We use a list of 2170 domains from Lasser, Aroyehun, Almog Simchon, *et al.* [147] labeled as “unreliable.” The authors compiled this list using ratings combining 9 academic and professional fact-checking sources. These ratings have high agreement with other lists of unreliable domains including proprietary news rating service NewsGuard (Krippendorff’s $\alpha = .84$), but unlike NewsGuard, the lists are available publicly and for free. The full list is available on OSF: <https://osf.io/68mn9>.

3.2.2 Survey Experiments

Procedure

We ran two survey experiments on Lucid to assess the impact of exposure to vaccine (mis)information on future intentions to take a COVID vaccine. We did not explicitly exclude Study 1 participants from participating in Study 2 because of limitations of Lucid, the online platform we used run our study. However, because Lucid has an extremely large subject pool (greater than 300 million respondents according to their documentation) and draws participants across a wide set of survey providers, it is unlikely that participants were repeated from one study to another. Both studies ran using the following identical procedure. To reduce demand effects, we ran each study in the same Qualtrics survey as a separate, distractor study.

First, participants received information about the distractor study, and filled out demographic information and pre-treatment attitudes related to the other study. Then, they answered the following questions about their pre-treatment vaccination attitudes (exact survey flow can be found on OSF: <https://osf.io/68mn9/>).

The participants then completed the distractor study. After finishing, they were shown a screen saying that the first study was complete, and given instructions to turn on their audio for the second and final study. Participants were then randomized to see either a single piece of control content or treatment content (described below). Participants saw the headline of the story accompanied by a picture or video (if applicable), in the same style as one might see on social media. They could play the video, but we did not enforce viewership. Participants were asked if they would like to share the content on social media (Yes or No), and could not advance to the next screen for at least 15 seconds. After the exposure period, participants advanced to the next screen where they were asked their post-treatment vaccination attitudes.

Finally, all participants in the treatment condition were debriefed and, if they were exposed to misinformation, told they had been shown information debunked by fact-checkers. They were then told vaccines were safe and effective, and given links to the CDC about the benefits of vaccination.

Sample

We conducted both experiments on the survey platform Lucid. Although participants on Lucid have been shown to have lower attention than other survey platforms, we believe this lack of attentiveness is a benefit for our study’s purposes, since our goal is to generalize these

in-survey results to a social media context, in which users are similarly likely to have low baseline attentiveness.

Study 1 We sampled 12,222 participants on Lucid, quota-sampled to match the US distribution of age, gender, ethnicity, and geographical region. We prevented participants who failed two trivial attention checks from entering the survey, and additionally excluded participants who failed to complete the survey, leaving us with 8,603 participants (8,500 were pre-registered). Data collection ran from 3/17/22 - 5/22/2022. The sample was 50.4% female, and had an average age of 47.5 years. A balance check found that our sample was balanced on pre-treatment covariates, and we found no evidence of differential attrition (see B.0.4).

Study 2 We sampled 13,547 participants on Lucid, quota-sampled to match the US distribution of age, gender, ethnicity, and geographical region. Data collection ran from 7/14/2022-8/3/2022. As in Study 2, we prevented participants who failed two trivial attention checks from entering the survey, and additionally excluded participants who failed to complete the survey, leaving us with 10,122 participants (10,000 were pre-registered). The sample was 52.1% female, and had an average age of 47.2 years. A balance check found that our sample was balanced on pre-treatment covariates (see Section B.0.3). However, due to a rendering error in our control group we had a small but significant amount of attrition in the control group compared to the treatment group. Nonetheless, analyses show that this control-group attrition appears to be random, and a robustness check using Manski extreme bounds shows that it has no bearing on our substantive results (See B.0.4).

Stimuli

Study 1 We collected 40 pieces of vaccine-related misinformation that had previously been debunked by fact-checkers. Stimuli included posts and videos from social media sites, links to news stories from mainstream and low-quality outlets, and news video clips. We also collected 10 control items from the same mix of platforms, giving us 50 items overall. The full list of stimuli can be found on our OSF site: <https://osf.io/68mn9/>.

Study 2 Unlike Study 2, which selected content that was already debunked by fact-checkers, we chose to gather a more representative sample of URLs popular on Facebook. Using the CrowdTangle API, we pulled the 500 vaccine-related URLs with the highest number of interactions from 1/1/2022 to 4/26/2022 from both mainstream and low quality domains, respectively. We appended this set with an additional 21 popular Facebook URLs from mainstream domains that discussed side effects of the vaccine (as identified by RAs) to increase topical coverage. Our list of mainstream domains was adapted from Pennycook and Rand (2020)([54]) and our list of low-quality domains was adapted from the Iffy index (which is a subset of the low-quality domain list from [147] which we use to classify our Facebook URLs.) The full list of domains can be found on our OSF site: <https://osf.io/68mn9/>. We then filtered out URLs that were irrelevant, redundant, or out-of-date, leaving us with 191 candidate URLs. We identified 6 major topics from this set (vaccine mandates, boosters,

vaccine approval/safety, children’s vaccine, and vaccine side effects, and other) and randomly sampled an equal number of URLs for each topic, balanced across domain type (mainstream vs. low-quality). This procedure produced with a dataset of URLs stratified on domain type and topic, with 45 URLs from mainstream domains and 45 URLs from low-credibility domains. In addition to these 90 vaccine-related URLs, we gathered 10 control URLs that were entertainment and news URLs not related to the vaccine.

Content Ratings

Crowd Ratings For studies 1 and 2, we used CloudResearch’s Amazon Mechanical Turk platform to solicit labels about the extent to which each post or article was harmful to a person’s health. Each post was labeled by on average 23 raters. Additionally, we used the platform Lucid to gather crowd-ratings for each stimulus on the following dimensions: surprising, plausible, favorable to Democrats (vs. Republicans), and familiar.

Each item received on average 15 ratings per question for Study 1, and 17 ratings per question in Study 2. The full list URLs and their associated labels and a copy of the survey containing the exact wording of our questions can be found on our OSF site <https://osf.io/68mn9/>.

Expert Ratings We also had 2 professional fact-checkers vet all 90 the headlines and descriptions from Study 2. The percent agreement between the 2 fact-checkers was 79%. When the fact-checkers disagreed, we gave them the opportunity to change their rating in order to reach consensus. Of the 90 headlines, 25 were fact-checked as false or “potentially misleading” by both fact-checkers, 19 were fact-checked misleading by one fact-checker, and 46 were fact-checked as true. We labeled the URL “misinformation” if it had been fact-checked as false or misleading by both fact-checkers, consistent with Facebook’s rules for aggregating fact-checker ratings.

Ethics

Participants gave informed consent and were told that they might encounter false information as part of the task. After the study was completed, participants who had been exposed to previously debunked misinformation were debriefed and told the content they saw was false and debunked by fact-checkers. All participants in the treatment group who saw any vaccine information (true or false) were given accurate information about the safety and efficacy of the vaccine and directed to the CDC website for more information. All experimental studies and crowdsourcing tasks were reviewed by MIT’s IRB and deemed minimal risk and exempt (E-2443, E-4266, E-4195, E-4717).

Outcome Variables

Our primary outcome variable was a vaccine-intention index composed of four questions ranging from 0 (definitely would **NOT** take a vaccine) to 100 (definitely would take a vaccine). Because our experiment ran in 2022 after the initial rollout of the COVID-19 vaccine, we asked each participant a question about willingness to get a hypothetical future booster

dose of a COVID-19 vaccine. In addition, we asked about intentions to take a first dose (if the participant had not yet received a first dose), booster vaccination intentions (if the participant had not yet received a booster), and intentions to vaccinate a child (if the child had not been vaccinated). The vaccine index was calculated as the average of the four outcomes (where available) in order to increase power. See [B.0.1](#) for exact wording.

Analysis Procedure

Study 1: Estimating Causal Impact and Stimulus-Level Heterogeneity We fit a linear mixed effects model using the `lmer` package in R with our vaccine index as the dependent variable; a treatment dummy variable for whether or not the participant was exposed to vaccine misinformation or a neutral control; random slopes for treatment for each stimuli; and controls for gender, age, political leaning, and pre-treatment vaccination intentions. We also included an interaction term for pre-treatment vaccination intentions and our treatment variable. The formal model can be found in [B.0.11](#). To measure the causal impact of misinformation, our quantity-of-interest was the coefficient on the treatment dummy variable, corresponding to the average treatment effect (ATE) of vaccine misinformation on vaccination intentions. To measure the stimulus-level heterogeneity, our quantity of interest was the standard deviation of the stimulus-level random effects.

Study 2: Predicting Moderators of Treatment Effect We use the same methodology applied in [\[148\]](#), which examined heterogeneity in political ad treatment effects. We use a two-stage process, rather than a single multi-level model, because our content-level features (e.g. the extent to which the item implied the vaccine was harmful) only applied to treatment and not control stimuli. In the first stage, we estimate the effect of each treatment compared to the control group. Then, in the second stage, we predict variation across treatment effects using content-level features as predictors.

More specifically, in the first stage, we fit two separate models for studies 1 and 2, respectively, estimating the treatment effect of each stimulus on our vaccine index. We estimated these treatment effects using OLS with HC2 robust standard errors, with controls for pre-treatment vaccination intentions, gender, political leaning, and age. Control stimuli were given the same stimulus ID, and served as the reference group for the other stimuli.

Then in Stage 2, using the `metafor` package in R, we fit a hierarchical meta-regression with the treatment effects that we fit in Stage 1 as the dependent variable, content-level features as the regressors, and nested random effects for study and stimulus ID. We accounted for the fact that these treatment effects were correlated because of a common control group by using the block-diagonal variance-covariance matrix estimated in Stage 1 in our meta-regressions [\[149\]](#). The formal model is specified in [B.0.2](#).

We ran separate meta-regressions for each potential moderator – i) harmful-to-health, ii) surprising, iii) plausible, iv) favorable to Democrats vs. Republicans, and v) familiar, and a joint model with all moderators together. The meta-regression coefficients on the “harmful-to-health” variable is our main quantity-of-interest reported in the main text. The full result of this model and other supplementary models can be found in [B.0.5](#).

3.2.3 Facebook URL Predicted Treatment Effects

An overview of the pipeline used to predict treatment effects for our 13,206 Facebook URLs can be found in Figure 3.1. For more detail see the following sections.

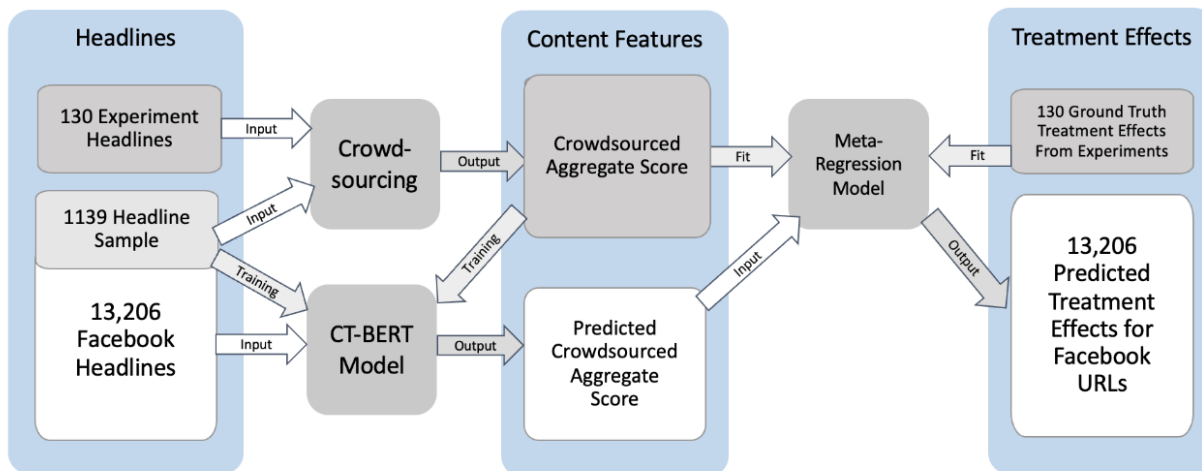


Figure 3.1: High-Level Overview of Treatment Effect Prediction

Crowdsourcing

We solicited crowdsourced judgments from lay people recruited from CloudResearch’s Amazon Mechanical Turk platform regarding whether each of the 130 vaccine-related items from Studies 1 and 2 would likely increase or decrease people’s willingness to get a Covid vaccine. In addition, we collected their judgments of the harmful-to-health rating and accuracy rating for each item (for precise wording, see B.0.17). We showed participants the headline and lead sentences of each article or social media post, as well as an image of the post as it would appear on social media. In total, we collected judgments from 177 laypeople. We excluded 7 participants who failed 2 trivial attention checks, and 22 participants who failed to complete the survey, leaving us with 148 participants total. Each participant rated 20 items, and each post received 22 responses on average. The sample was 47% female and had an average age of 39.6. We then created a “Crowdsourced Aggregate Score” variable by normalizing each of the three variables i) less vs. more willing to vaccinate, ii) harmful-to-health, and iii) accuracy and then averaging them together. We found that results with this aggregate score had better predicted performance than a model with the single “less vs. more willing to vaccinate” question alone and a model with the three features separately (see B.0.18 for comparison).

Meta-regression Model

To predict the treatment effects from the crowdsourced judgments, we fit a hierarchical meta-regression model with stimulus-level treatment effects as the dependent variable, the “Crowdsourced Aggregate Score” variable as the independent variable, and random effects

for treatment ID and experiment ID using the metafor package in R. This is the same analysis described in Methods Section 3.2.2 in which we predicted the treatment effect using content-level features, except we use the average crowd predictions instead of content-level descriptive features. The formal model is presented in B.0.11.

Predicting Features of Facebook URL Content

We train three different transformer-based models to predict the following dimensions of the Facebook URLs: i) less vs. more vax, ii) harmful-to-health, and iii) accuracy. We then average the results of these three models together to get a predicted “aggregate score,” which has a better performance than predicting the aggregate score directly. This improved performance is consistent with past machine learning research on ensemble models, which combine the predictions from multiple separate models and have shown to have better performance than a single model on its own [150].

Annotation Procedure

We sampled 1163 of the 13,206 Facebook URLs for labeling by the following procedure with the goal of oversampling URLs that were 1) covering a wide range of stories, 2) highly viewed, and 3) fact-checked misinformation. First, we randomly sampled one URL from each of the 500 story clusters described in Section 3.2.1. Then, we randomly sampled an additional 495 URLs, weighted by the public engagement data each URL received (as a proxy for the number of views, since the actual view data was not available outside of Meta’s environment). Finally, we took all 168 unique URLs fact-checked misinformation URLs, after filtering out duplicates and near duplicates of headlines. We took the union of these three sets, removing duplicates, leaving us with 1139 URLs.

Sampling Method	Count
Random Sample, by Story Cluster	500
Random Sample, by Public Engagement Received	495
All Fact Checked misinformation, filtering duplicated headlines	168
Total	1163
Total (after deduplicating identical URLs)	1139

Table 3.1: Labeled URL Data

We then used Amazon Mechanical Turk’s CloudResearch Platform to label the i) less-vs.-more-vax, ii) harmful-to-health, and iii) accuracy ratings of each URL using the same questions as described in Methods Section 3.2.3. The one change is that we only asked two questions in our accuracy battery – “accurate” and “biased” – to reduce the number of questions for the labellers. Each URL received on average 9 labels per headline.

Test-Train-Split

We split the 1169 URLs (the 1139 Facebook URLs, plus the 130 URLs from our two experiments) into train, validate, and test sets, stratifying on whether the crowd labeled the URL

as likely to increase or decrease vaccination. We first did a 15/85 split between our test set and training/validation set. We then did another 15/85 split into separate validation and training sets. This amounted to 176 URLs in our test set, 854 URLs in our training set, and 149 URLs in our validation set. For robustness, we also clustered the headlines by their embeddings, and performed a test/train/validate split on the cluster-level rather than the headline-level. We found similar performance to the original training procedure, which suggests the model would have good out-of-sample generalization properties. However, results from a model with this clustered training procedure were less conservative and had a higher false-positive rate (i.e. more vaccine content judged by the crowd to decrease vaccination intentions was labeled as increasing vaccination intentions), so we proceed with the original training procedure. We also trained a model that predicted the aggregate score directly, instead of separately predicting each component, and found it had very similar, but slightly worse performance. See [B.0.12](#) for results from these models.

Models

We fine-tuned a pre-trained COVID-Twitter-BERT model (CT-BERT) to predict each of our output variables [151]. This model showed a 10-30% improvement over the baseline BERT-large model on 5 specialized COVID-related datasets, including a vaccine sentiment task similar to the task we employ here. Furthermore, the model has been shown to have good performance on COVID-related fake-news detection tasks, consistently outperforming base-BERT and other model architectures [131].

Training Procedure We trained three separate models to predict our three different outcomes using Google Colab Pro. Each model was trained on the training set for 10 epochs and evaluated on the validation set. We selected the model with the best performance on the validation set as the final model. We used an AdamW optimizer with a learning rate of $2e-5$, max sequence length of 512 tokens (the max of CT-BERT), and a batch size of 4. All models were implemented using the HuggingFace transformers library.

Performance We present the performance of our model predicting the Crowdsourced Aggregate Score (Table 3.2) and a binary classification model predicting whether URL’s Crowdsourced Aggregate Score is below the scale midpoint of 3 (content which we deem as “skeptical”, i.e. likely to lower vaccination intentions, shown in Table (Table 3.3)). An analysis of the performance of the model alternative cutoffs for the aggregate score can be found in [B.0.7](#). For reference, the aggregate score model is measured on a 1-5 scale.

3.2.4 Predicting Treatment Effects for Facebook URLs

For each of the URLs, we input our aggregate score to the meta-regression model we fit in Methods Section 3.2.3 to get predictions of the treatment effect for each of the 13,206 Facebook URLs. Additionally, we parametrically bootstrap 1000 draws of our coefficients, giving us distributions of effects for each URL. We use these draws to compute confidence intervals and to visualize distributions for URL impact.

Metric	Value
rMSE	0.33
Correlation	0.87
MAE	0.25
Accuracy (within .5 of true value)	86%
Accuracy (within 1 of true value)	99%

Table 3.2: Performance for a model predicting the crowdsourced aggregate score of the Facebook URL

Metric	Value
Accuracy	91%
AUROC	97%
False Positive Rate	4%
True Positive Rate	80%

Table 3.3: Performance for a model predicting the binary skeptical vs. promoting/not skeptical rating of the Facebook URL

3.2.5 Combining Predicted Treatment Effects and Facebook Viewership Data

For each URL and draw of a treatment effect, we multiply the average predicted treatment effect by the number of views received by the URL in order to get an estimated impact for each URL. We then aggregate across the draws to get the overall distribution of impact across a set of URLs. We normalize this overall impact estimate by the total number of US Facebook users estimated to be on Facebook in Q1 2021 (233 million) [152]. Since content that substantially lowered vaccination intentions is the focus of our analysis, we filter to Facebook URLs that we classify as “skeptical,” defined as having a “Crowdsourced Aggregate Score” of less than 3, the scale midpoint. Our reason for choosing this cutoff is two-fold. First, our meta-regression model showed that headlines with a score less than 3 significantly lowered vaccination intentions (i.e. as shown in Figure 3.6, it is the point at which the upper 95% confidence interval crosses 0). Second, we find this cutoff has high accuracy (91%) and a low false positive rate (4%) on a binary classification task (see B.0.7 for cutoff analysis). Analysis of the full population of URLs and results for different thresholds can be found in B.0.15.

3.3 Results

3.3.1 Survey Experiments

Vaccine Misinformation Lowers Vaccination Intention

Our first experiment examines the question of whether misinformation lowers vaccination intentions on average, and if so, how much variation there is between stimuli. The study,

which ran on Lucid in March 2022 on 8603 participants, tested whether exposure to a single piece of vaccine misinformation drawn randomly from a set of 40 articles, videos, and posts previously debunked by fact-checkers lowered vaccination intentions compared to a neutral control. We estimate both the average treatment effect of misinformation and the distribution of potential treatment effects using a mixed effects model. All analyses are pre-registered.

Our results, summarized in Figure 3.2, support the conclusion that exposure to vaccine misinformation can reduce overall intentions to vaccinate. A single piece of vaccine misinformation decreased intentions to take a COVID-19 vaccine, measured by our COVID-19 vaccination index, by 1.5 percentage points ($p=.00004$). We also saw a significant decrease of similar magnitude for participants’ willingness to take a future hypothetical COVID-19 vaccine, and willingness to vaccinate their children. We did not find a significant impact on willingness to take a first dose of a COVID-19 vaccine, or willingness to take a booster. However, fewer participants in our sample were eligible to answer those questions (due to having already received a first dose of a vaccine or a booster, respectively), and thus those analyses are relatively less powered. Furthermore, in line with work finding mostly homogeneous effects of political persuasion between individuals [153] we also find no significant evidence of subject-level heterogeneity by pre-treatment vaccination intentions, gender, age, political party, or vaccine status. See Section B.0.6 for these analyses and other robustness checks.

Comparing across different stimuli, we find that not all misinformation is created equal. We find substantial variation in the treatment effects. Using a multi-level model fit with random slopes for treatment for each stimuli and our vaccination index as the outcome, we find that there is an overall standard deviation of .89 percentage points between stimuli. This constitutes approximately 60% of overall treatment effect, suggesting that the worst 10% of misinformation items had double the average effect – lowering vaccination intentions by 3% – and 10% of the stimuli had a treatment effect of 0. Simply because an item had been proven to be false did not mean that it lowered vaccination intentions. These results suggest that other dimensions of the content beyond veracity explain heterogeneity in the treatment effects. In the next section, we explore reasons for this variation in effect.

Vaccination intentions are lowered by implication of harm, not only falsity

While Study 1 demonstrated that misinformation lowered vaccination intentions on average, we found substantial variation in the effect of different stimuli. In a follow up study, we investigate explanations for this stimulus-level heterogeneity. We expand beyond misinformation and instead examine a representative sample of 90 vaccine-related articles sampled from Facebook using CrowdTangle, balanced across topic and domain quality. By testing a diverse set of content, we can discover which features predict vaccine hesitancy from the bottom-up.

One potential explanation for the variation is that content that emphasized potential health or death risks of the vaccine particularly decreased people’s willingness to get a vaccine by increasing their perceived risk of experiencing an adverse outcome. Prior research has found that concerns about vaccine side effects and safety were the most commonly cited reasons for hesitancy regarding the COVID-19 vaccine [128], [154]. Given this background,

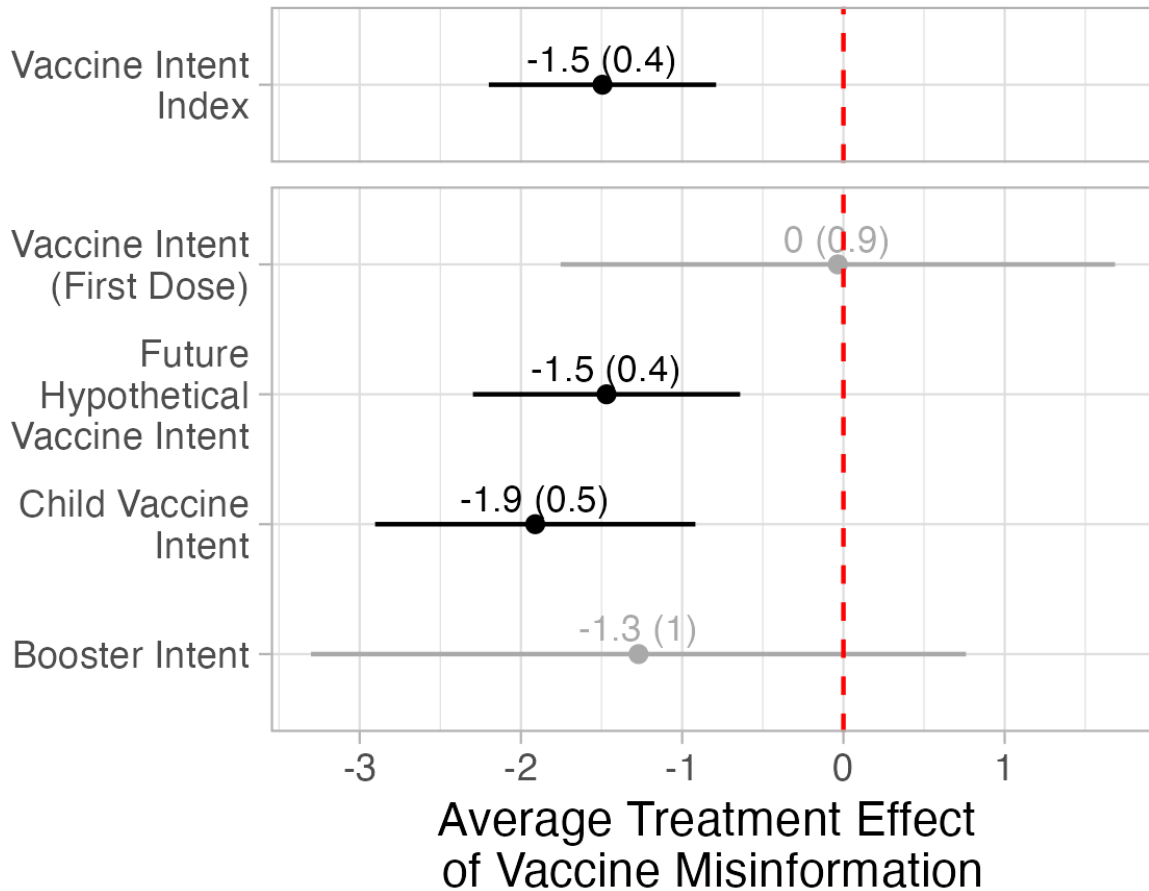


Figure 3.2: The effect of misinformation on intentions to take a vaccine from Study 1

and an exploratory analysis of our Study 1 data, we pre-registered a hypothesis that the extent to which an item suggested the vaccine was harmful to a person’s health would be associated with more negative treatment effects.

We recruited raters from the online platform Lucid to label the content in Studies 1 and 2 on whether it was 1) surprising, 2) plausible, 3) favorable to Democrats (vs. Republicans), 4) familiar and 5) whether the item suggested the vaccine was harmful (vs. helpful) to a person’s health. We then ran a random-effects meta-regression using the treatment effects of each headline on the overall vaccine index as our main outcome variable to determine whether any features consistently predicted the magnitude of the treatment effect.

Results from Study 1, Study 2, and their precision-weighted average can be found in Figure 3.3. We find support for our hypothesis that the extent to which an item suggests the vaccine is harmful to a person’s health predicts the treatment effect among a representative sample of vaccine content. In Study 2, we found a significant negative main effect for our harmful-to-health variable on our vaccination intention index. Stories that suggested the vaccine was 1 scale point more harmful-to-health were associated with a 0.5 (SE: .21, $p = .013$) percentage point (pp) decrease in our vaccination intention index. This means that an item rated as “Very Harmful” to a person’s health would have had an average marginal

effect of -1.2pp (95% CI: $-2.4, -.16$), compared to a $-.2\text{pp}$ (95% CI: $-1.0, .60$) for an item rated as “Neither harmful or helpful.” A multilevel-meta regression including both Study 1 and Study 2 studies demonstrated the robustness of the results, see Figure 3.4. Across both studies, increasing the harmful-to-health measure by 1 scale point was associated with an effect of -0.67pp (SE:0.18, $p = .0002$) for a model with just harmful-to-health as a predictor, and $-.50\text{pp}$, SE: 0.24, $p = .027$) for a model including other potential moderating variables.

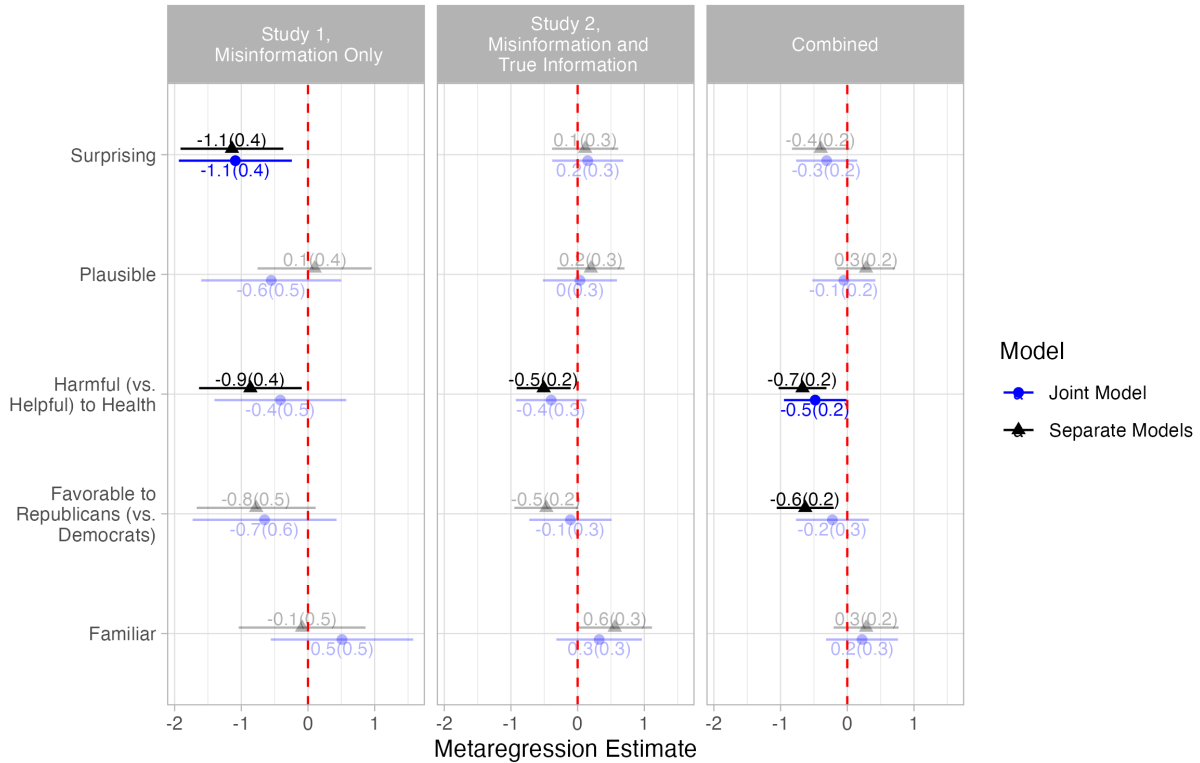


Figure 3.3: Coefficients from respective meta-regressions testing how different features of content moderate the treatment effects in i) Study 1, ii) Study 2, and a iii) multi-level meta-regression combining both studies. The black points show the results from a model testing all moderators separately, and the blue points are from a joint model that contains all 5 moderators.

No other content feature consistently explained variation in treatment effects across both studies across both studies, as can be seen in Figure 3.3. We also find no significant main effect of low-quality domain (vs. mainstream domain) on vaccination intentions ($= -0.26$, SE: .23, $p = .25$). These analyses, as well as a post-hoc exploration of additional potential moderators can be found in Section B.0.5. The only additional dimensions that significantly moderated the effect were measures of how harmful to one’s own health or children’s health the content was, bolstering the interpretation that questioning vaccine safety drives hesitancy.

Post-hoc, we also had 2 fact-checkers rate the veracity of all 90 items in Study 2. Replicating the findings in Study 1, we found that misinformation lowered vaccination intentions in this sample (ATE= -1.2 , $p=.035$). However, in a model that included both harmful-to-health

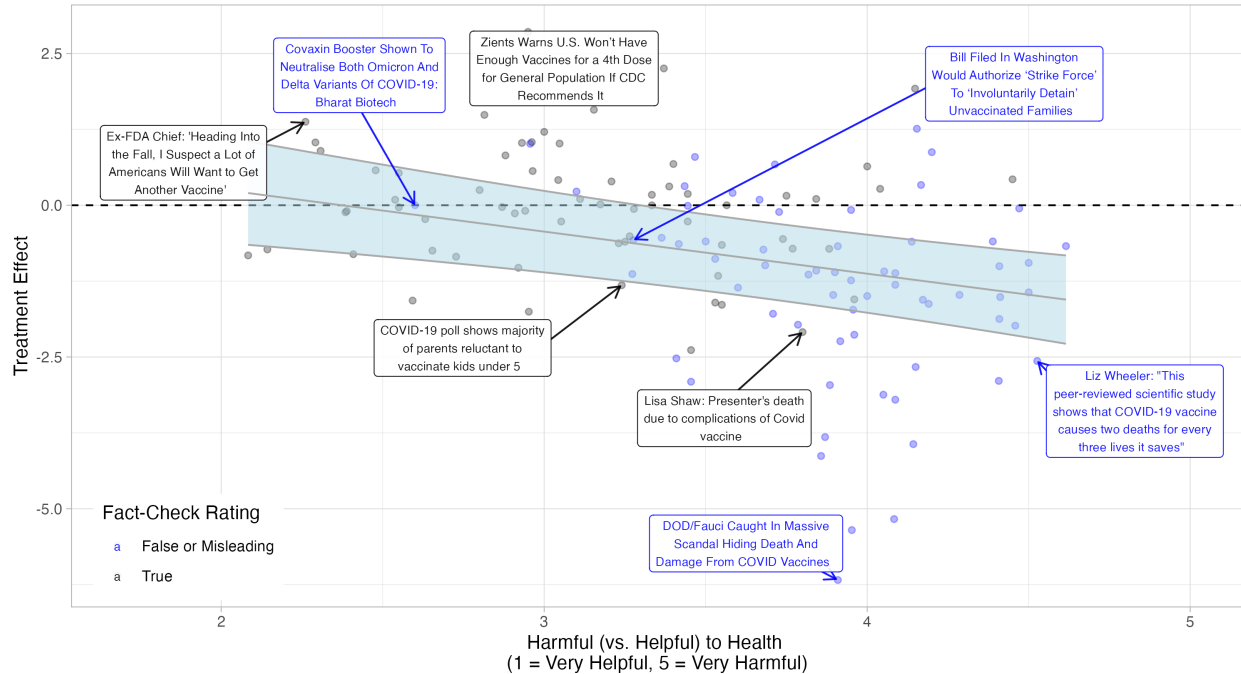


Figure 3.4: A scatter plot of all treatments across Studies 1 and 2 where each point represents a single treatment. The gray line is the best-fit line from the meta-regression model with harmful-to-health as the predictor, and the blue ribbon is the bootstrapped 95% confidence interval of the estimate. Misinformation is labeled in red, while factually-accurate content is labeled in black.

and veracity to predict the treatment effects across both studies, only harmful-to-health was significant ($B = -.38$, $p = .005$), while veracity was not ($B = -.21$, $p = .18$), and an interaction was also not significant ($B = -.19$, $p = .17$). These results indicate that suggesting that the vaccine caused harm reduced vaccination intentions regardless of whether or not the item was factually inaccurate.

3.3.2 Exposure on Facebook

In the previous section, we showed that stories that either claimed or implied that vaccines were harmful to health lowered vaccination intentions regardless of their veracity. Conversely, articles that did not imply harm could not be shown to lower vaccination intentions. We now move on to quantifying exposure to COVID-19 vaccine content on Facebook during the initial rollout of the COVID-19 vaccine in the US. Using Facebook’s URL Shares dataset, we identify 13,206 high-shared URLs about the COVID-19 vaccine from the first 3 months of 2021. This data contains view counts of each URLs, as well as information about whether Facebook referred the URL to third-party fact-checkers and if so, its rating. A more complete accounting of Facebook’s fact-checking procedure can be found in the Methods section. We find that misinformation content makes up a negligible percentage of viewership on Facebook. As presented in Figure 3.5, Panel A, URLs rated as false, out-of-context, or mixture – which we will refer to as “misinformation” in our subsequent analysis – received 8.7 million

views, accounting for only 0.3% of the over 3 billion views during this time period. These numbers were dwarfed by the 3 billion views received by URLs not sent to fact-checkers. An analysis that classified content by the quality of the parent domain, rather than the fact-check rating, yielded substantially similar results: Only 3.5% of views went to domains rated as low-credibility.

Thus, exposure to false vaccine content on Facebook was relatively infrequent, either due to user viewership preferences or explicit downranking by Facebook of misinformation content. These numbers might lead one to conclude that vaccine content on Facebook was unlikely to have caused vaccine hesitancy during the vaccine’s initial rollout. However, those conclusions are premature without a closer examination of unvetted content, which constitutes the vast majority of views in the dataset.

Figure 3.5, Panel B shows exposure to the top 10 most-viewed vaccine-related stories among all content. For interpretability, we cluster together stories with similar headlines covering the same event. An examination of these highly-visible stories reveals that several articles published by mainstream news organizations cast doubt on the safety and efficacy of the vaccine. For example, the most viewed URL across all 13,206 URLs during this time period is a Chicago Tribune article titled “A Healthy Doctor Died Two Weeks After Getting a COVID vaccine; CDC is investigating why.” This particular URL was seen by over 50 million Americans on Facebook – more than 20% of Facebook’s US user base, while URLs related to this story were viewed over 65 million times on Facebook, more than 7X the number of views on all misinformation combined. We emphasize that this story, and others like it, was factually accurate and in many cases indicated the uncertainty surrounding the true cause of death. Nonetheless, its clear implication was that the vaccine may be harmful to health, and thus may have had a substantial negative impact on vaccination intentions. Prior work has termed this ambiguous type of content that could – intentionally or not – lower vaccination intentions “vaccine-skeptical,” and hence we use the same definition going forward [155].

3.3.3 Predicting Treatment Effects for Facebook URLs

The Crowd Can Predict Variation in Treatment Effects

Thus far, we have found that content that suggests the vaccine was harmful to a person’s health 1) causally lowered vaccination intentions, and 2) gained widespread attention on Facebook. A natural next step is to try to estimate potential influence that these Facebook URLs had on vaccination intentions. In this section, we show that the wisdom of the crowd, calibrated to ground-truth estimates and augmented with natural language processing, can be used to predict the expected treatment effects remarkably well.

We asked 148 crowd raters recruited from Cloud Research’s Amazon Mechanical Turk panel to predict how vaccine skeptical vs. promoting each of the 130 treatments from Studies 1 and 2 were by asking them to guess whether a post would cause users to be more or less likely to take a COVID-19 vaccine on a 1 to 5 scale. We find that this crowd prediction of the treatment effect was correlated at $r=.40$ ($p<.00001$) with the treatment effects from the experiments. While this estimate is in line with other research showing layperson crowds can successfully predict variation treatment effects for nudge interventions about the flu vaccine [142], it is a lower bound of the true correlation due to sampling error in the individual

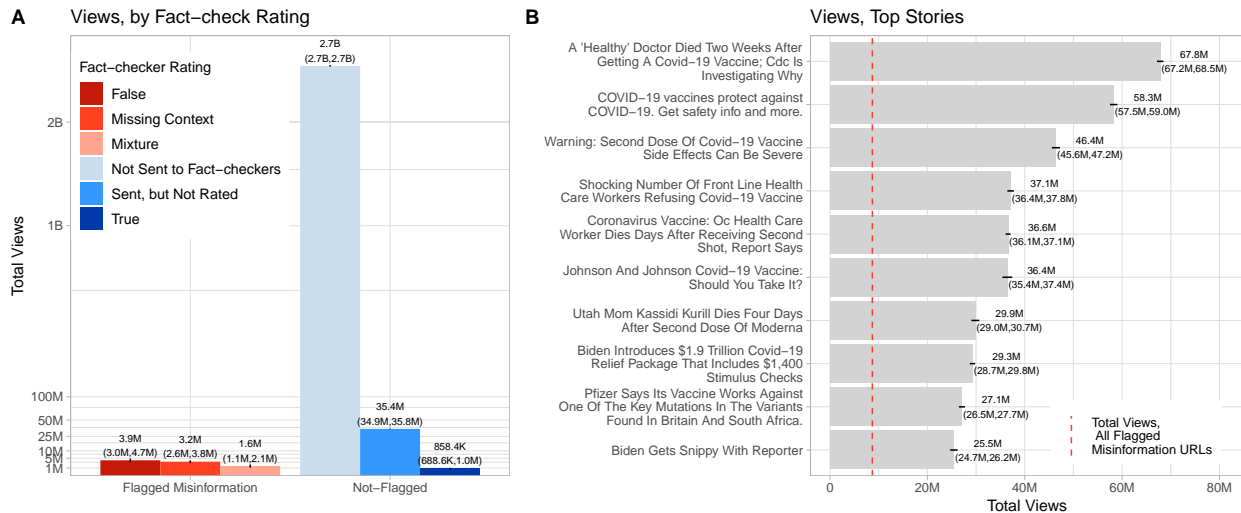


Figure 3.5: Exposure to vaccine related content on Facebook shared greater than 100 times on Facebook during the first 3 months of 2021. Panel A shows the total views for misinformation vs. non-misinformation content, broken down by fact-checker rating. The y-axis is square-root scaled for better visualization of misinformation content, which received only 0.3% of views during this time period. Panel B shows the total views of the top 10 most popular stories among all content. We cluster these URLs based on the tf-idf scores of their headlines and descriptions using the k-means algorithm. The aggregate number of views on all misinformation URLs is indicated by the red dashed line.

treatment effects. This sampling error is substantial because our study was designed to uncover generalizable features of content that cause vaccine hesitancy, rather than precisely estimate the effect of any single stimulus. Therefore, in order to assess the performance of the crowd while accounting the sampling error in our estimates, we run a random effects meta-regression model with our crowdsourced prediction as the regressor. We find that our crowdsourced skeptical vs. promoting prediction is a highly significant predictor of the treatment effects ($B=0.67$, $p=.0005$). Furthermore, we see that this model explains the majority of non-sampling variance in the results. The I^2 , the percent of residual heterogeneity not attributable to sampling variation, is 18.5%, suggesting that the model explains 81.5% of the non-sampling variation in treatment effects.

While these results are impressive, we can improve upon them by incorporating our results from Section II. We average together the crowdsourced skeptical vs. promoting prediction with crowdsourced predictions of accuracy (which agreed highly with expert judgments, $r = .72$, $p < .000001$) and our harmful-to-health variable to create an “Crowdsourced Aggregate Score.” This aggregate score significantly predicts the treatment effects ($B=.84$, $p=.0001$), and I^2 of 13.4%. These results are visualized in Figure 3.5. These results demonstrate that while the crowd might not predict a given individual treatment effect with high accuracy (due in part to the sampling error), it can successfully predict the expected average treatment effects across the range of crowdsourced predictions. Since we are ultimately interested in understanding the overall impact of Facebook content across thousands of headlines, rather than the precise impact of any single headline, these results show that crowdsourcing is a

capable tool for our task.

One thing to note is that while our results demonstrate that laypeople’s judgments can predict relative variation in the treatment effects, they make systematic errors when considering the magnitude. Examining the confidence interval of our meta-analytic prediction (shaded in gray), we see that while content that the crowd predicts will lower vaccination intentions (i.e. that is less than the scale midpoint of 3, which we label “skeptical”) does significantly lower vaccine intentions, the content that the crowd predicts will increase vaccination intentions (i.e. that is greater than the scale midpoint, which we label “not-skeptical/promoting”) has null effects on vaccination intentions. These results are in line with studies that have found an asymmetrical relationship between the ability to increase vs. decrease intentions to get a vaccine [156]. Our model adjusts for this overconfidence of the crowd by scaling the crowdsourced estimates to the actual range of treatment effects.

Natural-Language-Processing (NLP) Models Can Scale Crowd Efforts

Given the success of the crowd at predicting our ground-truth 130 treatment effects, we now apply the crowdsourcing method described in the section above to our 13,206 Facebook URLs. However, since gathering crowdsourced judgments for all URLs would have been too costly to procure, we solicited crowdsourced scores for only a subset of 1200 URLs and then trained a machine learning model to predict the crowdsourced scores for the entire population from the headlines and descriptions of the URLs. We find that our CT-BERT model is highly accurate at predicting the crowdsourced aggregate score; 84% of predicted aggregate scores were within half of a scale-point of the true aggregate score. We also use a cutoff of 3 to delineate “promoting” from “skeptical” content, since it is both the scale midpoint and the value at which the 95% confidence interval crosses 0 in our meta-analysis. On a binary classification task predicting whether the URL was vaccine-skeptical, the model had a 97% area-under-the-receiver-operating-curve (AUC), 91% accuracy, and a 4% false-positive-rate (i.e. only 4% of promoting URLs are incorrectly labeled as skeptical). Explorations of the results at other cutoffs can be found in Section B.0.7.

We pass these predicted aggregate scores to our meta-regression model to predict treatment effects for our URLs. While we predict treatment effects for the entire set of URLs, we will focus the majority of our analysis on “vaccine-skeptical” URLs, which significantly lowered intentions to vaccinate, rather than the “vaccine-promoting / not-skeptical” URLs, which had null effects.

3.3.4 Impact Estimates

Finally, we estimate the overall impact of vaccine content on Facebook during the first quarter of 2021. We present the respective distributions of these predicted treatment effects for the 186 fact-checked misinformation URLs and 13,020 factually-accurate URLs in Figure 3.7, Panel A. Conditional on viewership, the typical misinformation URL is substantially more likely to lower vaccination intentions than the typical factually-accurate URL. The median misinformation URL has a predicted treatment effect of -1.34pp (95% QI: [-1.91,-.766]), more than four times the effect of a typical factually-accurate URL of -0.3pp (95% QI: [-.91,.31]). However, in Figure 3.7, Panel B. A we see that when the treatment estimates are weighted by

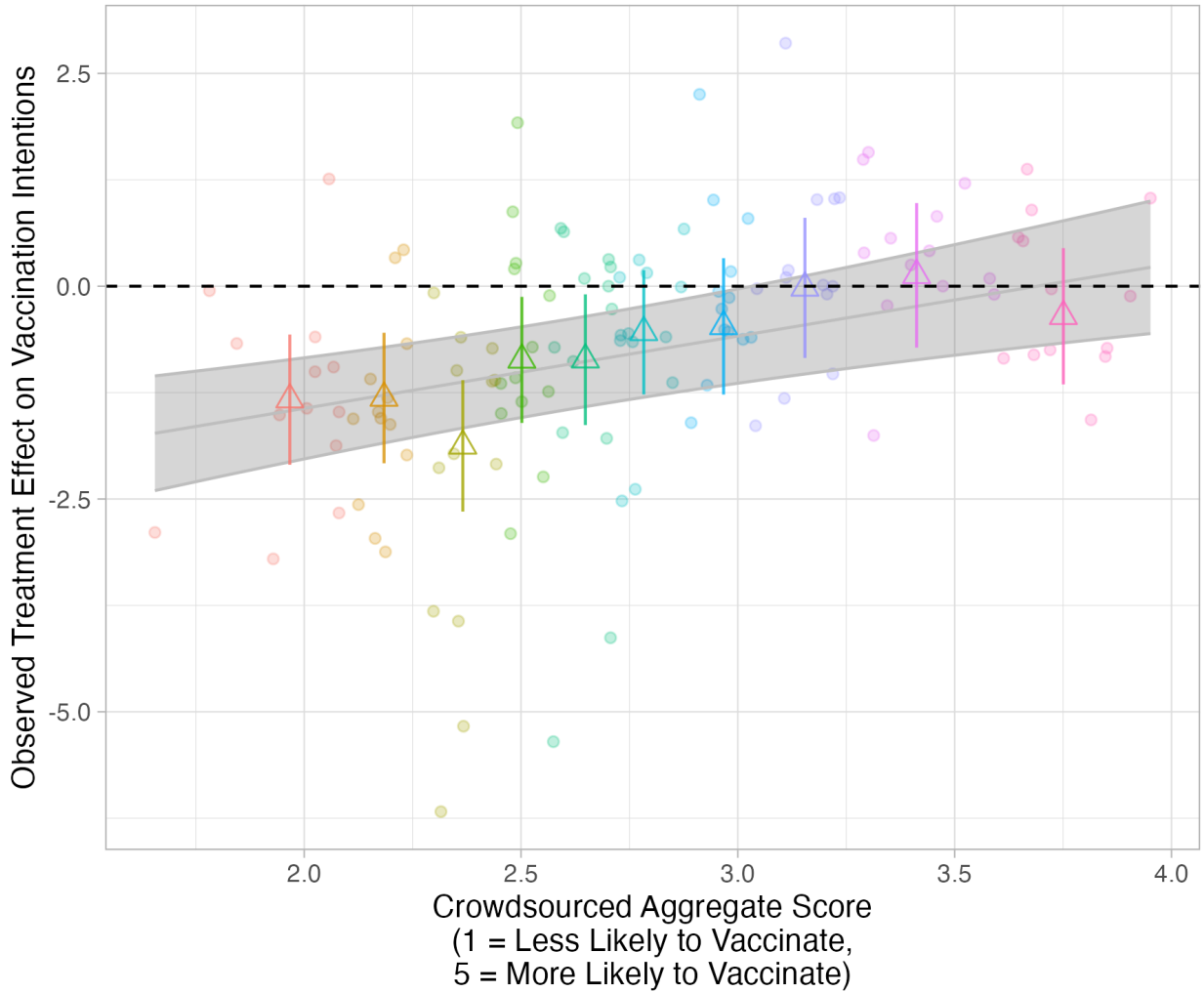


Figure 3.6: Treatment effect on vaccination intentions as a function of the Crowdscore Aggregate Score. Each point corresponds to one of the 130 items in Studies 1 and 2 and are colored by decile. The overlaid gray line is the best-fit line and 95% confidence-interval from a random effects meta-regression with treatment effect as the outcome variable, the crowdsourced score as a moderator, and random effects for item and experiment. Each colored-triangle shows the meta-analytic average within each decile of the crowdsourced score and shows that the results are not dependent on the linearity assumption.

viewership, the relative impact of these misinformation URLs was negligible. Many factually-accurate URLs were predicted to decrease vaccination intentions at magnitudes comparable to misinformation URLs – and with much greater viewership.

To estimate the impact of “vaccine-skeptical” content, we take the product of these estimated treated effects and the exposure data from Section 1 for each URL posted on Facebook. For interpretability, we normalize this overall impact estimate by the total number of US Facebook users (approximately 230 million) [152]. These results, shown in Figure 3.8 Panel A, show that this factually-accurate vaccine-skeptical content had a much larger negative

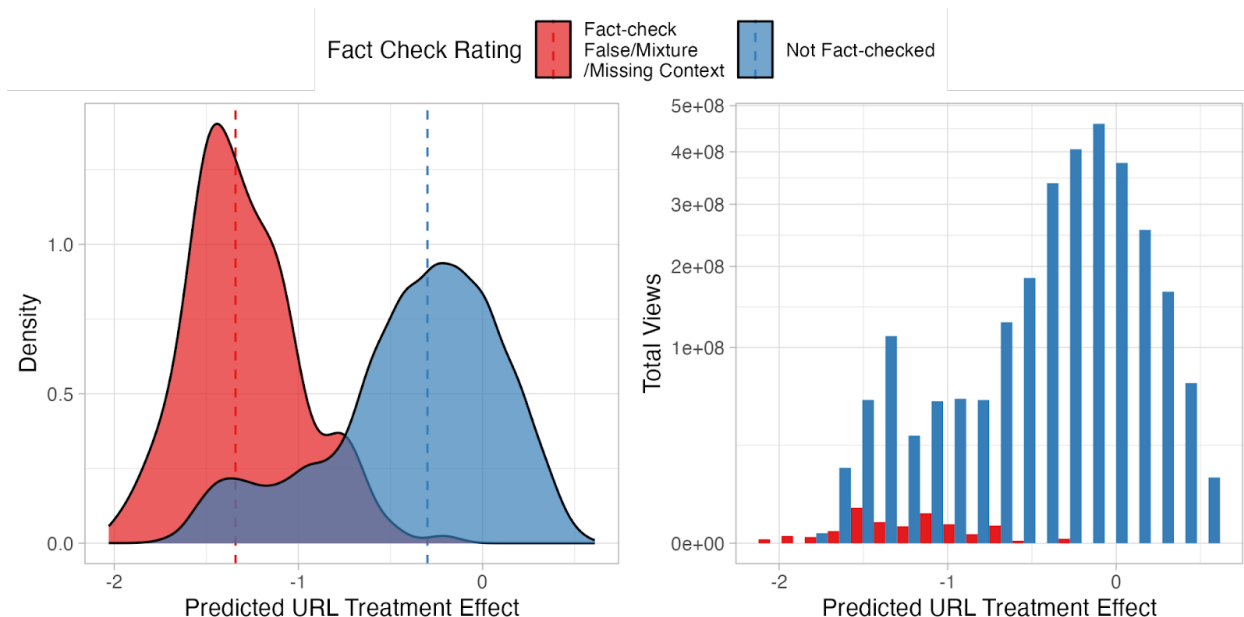


Figure 3.7: Distribution of predicted URL treatment effect on vaccination intentions for misinformation (shown in red) vs. factually-accurate content (shown in blue). Panel A shows the density plots for predicted treatment effects. Dashed lines represent the medians of the distributions. Panel B shows the same histogram of URL treatment effects, weighted by number of views each URL received. Note that the y-axis is shown on a square-root scale for better visualization of the misinformation.

overall impact per user than misinformation content. We estimate that factually-accurate skeptical content lowered vaccination rates by -2.28pp (CI: $-3.4, -.99$) per US Facebook user, compared to -0.05pp (CI: $-.07, -.05$) for misinformation – an almost 50-fold difference. This difference was driven almost entirely by exposure. Factually-accurate content accounted for 98% of the over 500 million views of vaccine-skeptical content.

What type of content drove this outsized impact? Figure 3.8 Panel B examines the predicted most harmful stories among misinformation vs. factually-accurate content, respectively. Among the factually accurate content, we can see that coverage of “healthy” people’s deaths following the vaccine reported from mainstream or local news outlets gained significant traction on Facebook. The most impactful misinformation story was only 1.5% as impactful as the most impactful factual story, which was the Chicago Tribune story about the “healthy” Miami doctor dying post-vaccine described earlier in Section 1. The other top factually-accurate stories show that this “healthy doctor” story was not a one-off occurrence. Coverage of young, healthy people’s deaths following the vaccine received disproportionate reach, and therefore had disproportionate impact, during this time period.

What sources publish this vaccine-skeptical content? Contrary to popular belief, high credibility domains drove more vaccine hesitancy than low credibility domains, even though a much greater proportion of content from low-credibility domains was vaccine-skeptical than was promoting. Figure 3.9 Panel A shows the number of vaccine skeptical vs. promoting URLs for low vs. high quality credibility, respectively. Two-out-of-three of vaccine

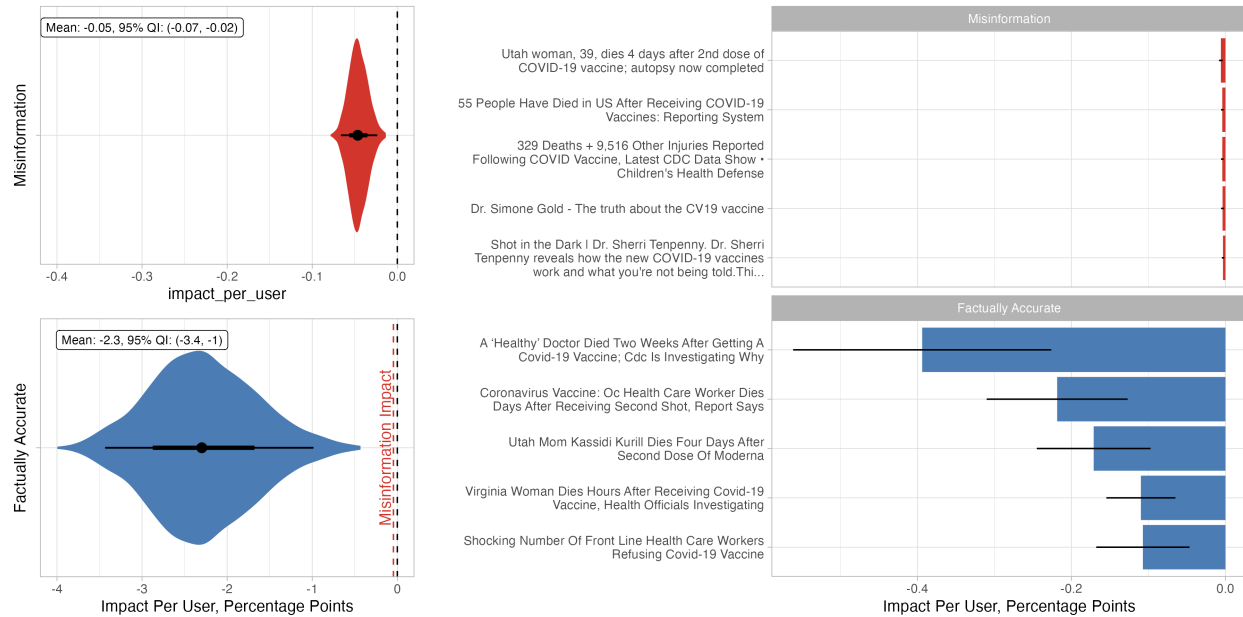


Figure 3.8: Impact of vaccine-skeptical URLs for fact-checked misinformation vs. not misinformation, respectively. Panel A shows the distribution of total impact across all vaccine-skeptical factually-accurate and misinformation URLs, normalized by the number of US Facebook users. Estimates are shown with 50 and 95% quantile intervals, calculated from a parametric bootstrap of our coefficients. Note that the scales for misinformation differ from factually accurate information; we label the average misinformation impact with a red dashed line for reference. Panel B shows the relative impact of the top most influential events for factually-accurate and misinformation URLs, with the most influential event normalized to -1. Error bars are 95% Quantile Intervals from our parametric bootstrap of our model coefficients.

URLs published by low-quality outlets were vaccine skeptical, compared to one-out-of-five of vaccine URLs by high-quality outlets. However, high-credibility domains still published approximately 1.6X more vaccine-skeptical URLs overall. And, as Figure 3.9 Panel B shows, these high-credibility domain URLs had much greater reach, and thus, had a much larger overall impact. Low credibility domains were only responsible for 9% of the total negative impact on vaccination intentions. A list of the top harmful domains can be found in Section B.0.9. Despite the worries about “misinformation superspreaders” and the “Disinformation Dozen”, mainstream news outlets like the New York Post and Fox News, as well as local news outlets, dominate the list.

3.4 Discussion

Our analysis answers long-standing questions about the effect of social media on large-scale societal outcomes. We estimate that vaccine-skeptical content on Facebook did plausibly lower US vaccination intentions by approximately 2.3 percentage points per Facebook user. However, contrary to popular wisdom, we show this effect was driven predominantly by

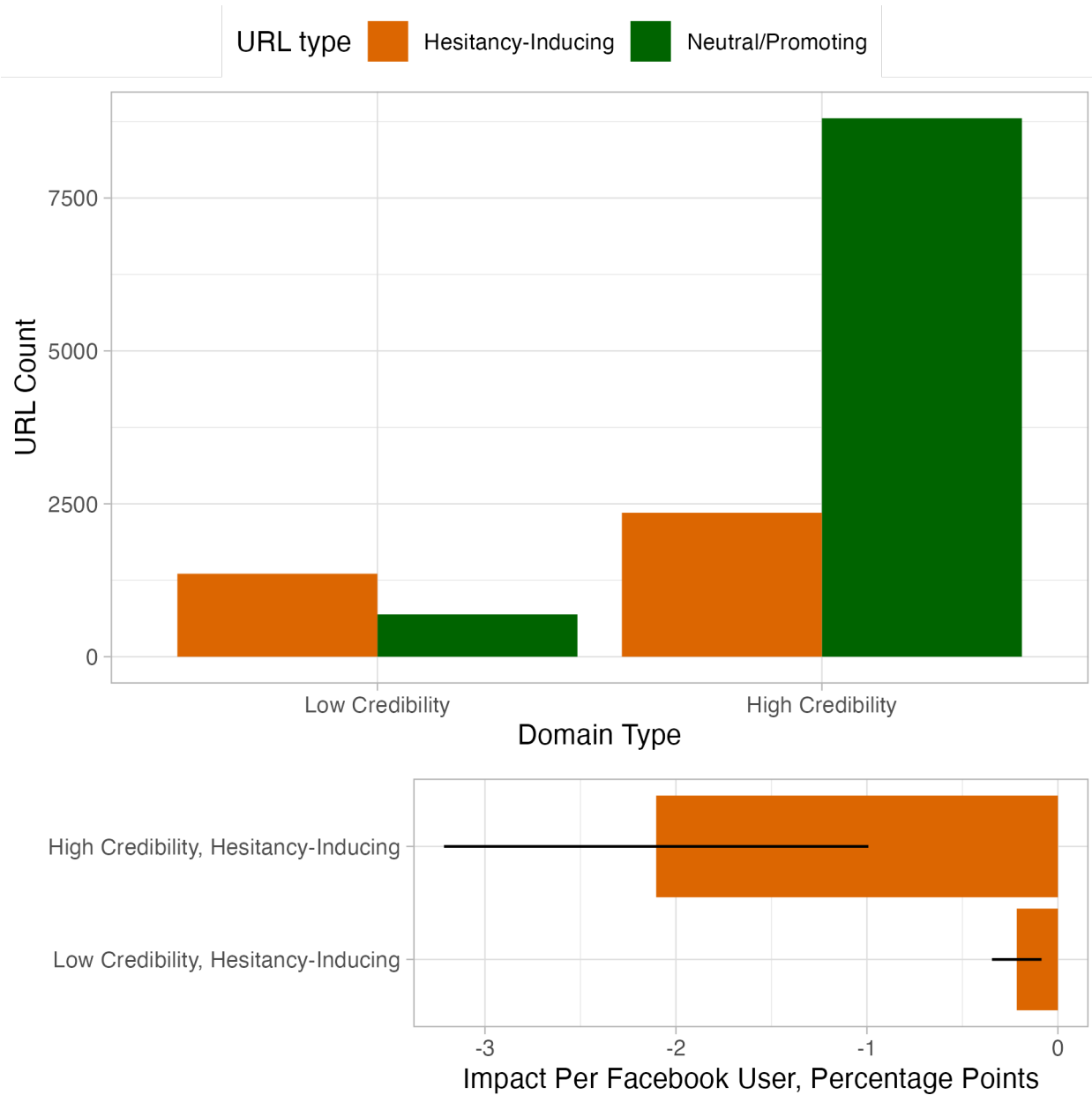


Figure 3.9: Comparison of low quality and high quality domains. Panel A shows the number of vaccine skeptical (i.e. hesitancy-inducing) vs. neutral/promoting URLs from low credibility vs. high credibility domains, respectively. Panel B shows the percentage of impact (measures as the product of exposure to vaccine-skeptical URLs and the respective predicted treatment effect of each URL) attributed to low-credibility vs. high credibility domains. We use the term hesitancy-incuding interchangeably with the term vaccine-skeptical

vaccine-skeptical content from mainstream sites, rather than false vaccine conspiracies published by fringe outlets.

These findings allow us to re-evaluate the efficacy of the most common interventions for identifying and fighting misinformation as tools for preventing harm. The typical ap-

proaches identify misinformation using third-party fact-checker labels or ratings of domain quality. They involve strategies like surfacing fact-checker labels or corrections, penalizing low-quality domains, or scaling digital literacy interventions that advise “checking the source” of content [108], [157]–[162]. Even automated systems designed to detect and limit the spread of fake news online primarily use databases of fact-checked claims as training data [31], [135], [163]. While we cannot know what a counterfactual world would look like without these interventions, our analysis shows that content targeted by these interventions represented only a small proportion of the potentially harmful vaccine content on social media. We cannot discern the reason for these low rates of exposure; it could be that the interventions were successful or social media users simply preferred other content. Given that misinformation was substantially harmful, conditional on exposure, any efforts by Facebook to limit its spread likely improved vaccination rates. However, none of the interventions mentioned would have prevented the spread of the type of content identified as most negatively impactful by our models: factually accurate but nonetheless vaccine-skeptical stories published by mainstream outlets including the Pulitzer-prize winning Chicago Tribune.

Instead, our results suggest that interventions should devote more attention on the reach and harmful influence of content, in addition to veracity. The information ecosystem might be vast, but human attention is finite. Improving the quality of highly viewed content, much of which comes from popular influencers or the mainstream media, is likely a more efficient strategy for improving people’s information diets than playing whack-a-mole against an ever-growing, but little seen universe of false content. Mainstream media outlets with widespread reach should consider that in spite of the caveats and acknowledgment of uncertainty included in their coverage, readers might respond in ways that cause real-world harm, especially in a social media environment where context is lost. Rather than focusing exclusively on the accuracy of the facts they report, journalists should consider whether the resulting stories will leave readers with an accurate worldview.

Our results also emphasize the need for researchers to devote more attention to understanding and tracking harmful content, irrespective of veracity. However, studying harm comes with its own challenges. First, identifying the causal mechanisms driving social ills is difficult. Second, even once identified, harmful content is hard to track at the scale of social media. While not a panacea, our work offers a framework for addressing both of these challenges.

First, we offer a methodology for discovering which content causes harm from the “bottom-up,” rather than relying on the (potentially biased) inclinations of researchers. This work addresses the “stimulus-as-fixed-effect” fallacy, a common threat to generalizability in social science research, and contributes to the growing literature on doing causal inference using latent-dimensions of treatments in large scale social data [134], [164]. By analyzing a large, representative set of content, we are able to identify which features of content cause vaccine hesitancy in a way that is generalizable to the stimulus population of interest (in our case, popular Facebook vaccine stories). While we apply this methodology to study vaccine hesitancy, we imagine that a similar “bottom-up” methodology could be used to identify the drivers of other potential harm outcomes, like political polarization or support for undemocratic practices.

Second, in regards to the challenge of tracking harmful content, we show that crowdsourcing and natural-language-processing techniques, calibrated to the distribution of ground-

truth treatment effects, are a promising solution to tracking harmful content at scale. Although, as past research has shown, laypeople often overstate the magnitude and direction of individual treatment effects [142], [165], here we show that they are able to predict variation across treatment effects with high accuracy. These judgments, coupled with NLP methods, can then be used to predict treatment effects for new samples of content. This work contributes to the large body of literature showing the power of the wisdom of the crowds across a variety of fields, and in particular, the power of crowd-machine hybrid models [18], [163], [166], [167]. While more work is needed to assess whether the crowd can predict treatment effects for other topics beyond vaccines, if the crowd’s performance generalizes, these methods could provide a way to predict the persuasive effects of content at scale without the need to run a large number of expensive, slow, and often underpowered RCTs.

Although our work offers novel contributions to the literature, it nonetheless has limitations. One drawback is that our experiments and our observational data come from different time periods. The Facebook viewership data (which is available only at a many month lag) is from the first quarter of 2021, whereas we ran our experiments in mid 2022. Ideally, this testing would have happened at the same time. To analyze whether results would have substantially differed due to timing, we perform several robustness checks. First, we analyzed the most impactful content from our experiment of 90 items, using the number of Facebook interactions as a proxy for exposure, and found similar patterns. For example, the most negatively impactful headline in our experimental dataset was an article published by the BBC with the headline “Lisa Shaw: Presenter’s death due to complications of Covid vaccine.”

Second, we consider how the persuasive effects might have differed if we had measured them in early 2021 instead of 2022. A Bayesian account of persuasion suggests that people might have been less set in their beliefs about vaccination during the initial vaccine rollout than during our testing period, and thus, promoting content might have had a positive rather than null impact on vaccination intentions during this time. In Section B.0.10, we explore what overall impact estimates would have been assuming different, non-null, average treatment effects for promoting content. Using as a benchmark the average treatment effect from experiments that tested the impact of promoting content on Facebook on willingness to get a vaccine in early 2021, we calculate the net impact of promoting and skeptical vaccine content on Facebook would have been -1.44pp per user, rather than -2.3pp. That being said, there is reason to believe our estimate of -2.33pp per user decrease for vaccine-skeptical content is an underestimate as well. An early 2021 experiment found that vaccine misinformation lowered vaccination intentions by 6% – a much larger magnitude, perhaps indicative of greater persuadability. Different assumptions could lead to different conclusions, and we by no means consider our analysis the final word on Facebook’s impact on vaccination rates. Our major contribution is a framework for evaluating impact, rather than a single definitive number.

Another potential limitation is that while our experiment participants were randomly exposed to content in a survey context, vaccine-hesitant users on Facebook might have actively sought out vaccine-skeptical content or been selectively targeted to see it by Facebook’s algorithm. Although we do not find significant evidence of participant-level treatment effect heterogeneity, we cannot rule out the possibility that exposure to vaccine skeptical content was concentrated in users who were likely going to refuse the vaccine anyway. To this end, we analyze how the concentration of vaccine skeptical information differed among different de-

mographic populations in Section B.0.16. As one might expect, very conservative users had information diets composed of the greatest proportion of vaccine-skeptical content (27%). However, all political groups saw at least 10% vaccine-skeptical content, and perhaps most concerningly, 23% of content viewed by non-political Facebook users was vaccine-skeptical (21% from high-credibility sources). While we cannot conclude with certainty that vaccine skeptical content was not concentrated even within these demographic buckets, the fact that over 20% of Facebook’s US population viewed the Chicago Tribune “healthy doctor dies” story suggests that this vaccine-skeptical content achieved broad popularity in at least some cases. Nonetheless, understanding how influence might differ for users with different prior beliefs about vaccines, and how repeated exposure to vaccine-skeptical messages might change cumulative impact is an area for future research.

One other potential concern raised by our work is that efforts to mitigate harm by targeting ambiguous in addition to outright false content might curb freedom of expression. Content that is not a priori objectionable can cause objectionable outcomes. We do not assume there is a straightforward solution to this issue. However, we believe that no informed discussion of tradeoffs can happen without quantifying potential costs and benefits. In the context of public health, these costs could be substantial. In a counterfactual world in which people were not exposed to the most harmful vaccine content, our model suggests that vaccination rates would have been 2.3% higher among Facebook’s 233 million US users, translating to approximately 5 million more vaccinated Americans. Assuming that 248 additional vaccinations translate into an additional life saved as estimated in [168], this would imply that many lives could have been saved if this content had not been published or allowed to spread unchecked.

Appendix A

Chapter 2 Appendix

A.1 Model Robustness

A.1.1 Alternate Evaluation Metrics

Misleadingness Classification

We also evaluate an RF model predicting the misleadingness classification of tweets using F1 score and Accuracy as alternate metrics. Figure [A.1](#) shows these results. According to both metrics, "Content features" are the worst performing. Accuracy follows the same pattern as AUC, with "Political Features" as second worst, followed by "All Features", and then "Context Features" being the best. However, on F1 score, the "All Features" is second worst, followed by "Context", and then "Political" being the best.

Helpfulness Classification

We also evaluate an RF model predicting the helpfulness classification of notes using F1 score and Accuracy as alternate metrics. Figure [A.2](#) shows these results. Both metrics follow the same pattern as AUC, with "Content Features" being worse, then "Political Features", "Context Features", and "All Features", and being the best.

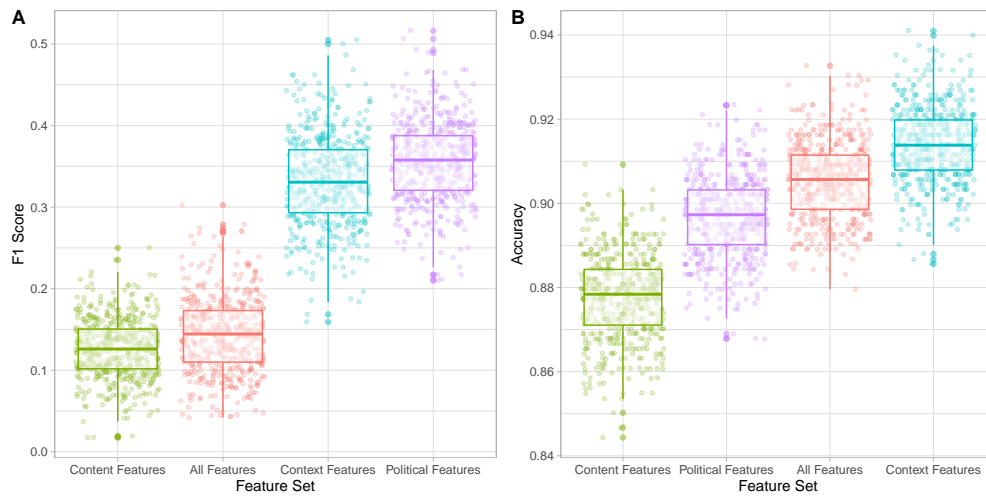


Figure A.1: A comparison of RF models predicting misleadingless classification of tweets, trained with different sets of feature using (A) F1 Score and (B) Accuracy as Evaluation Metrics

Panel A shows a boxplot showing the performance of different models using F1 score as a metric, Panel B shows a boxplot showing the performance of different models using Accuracy as the evaluation metric

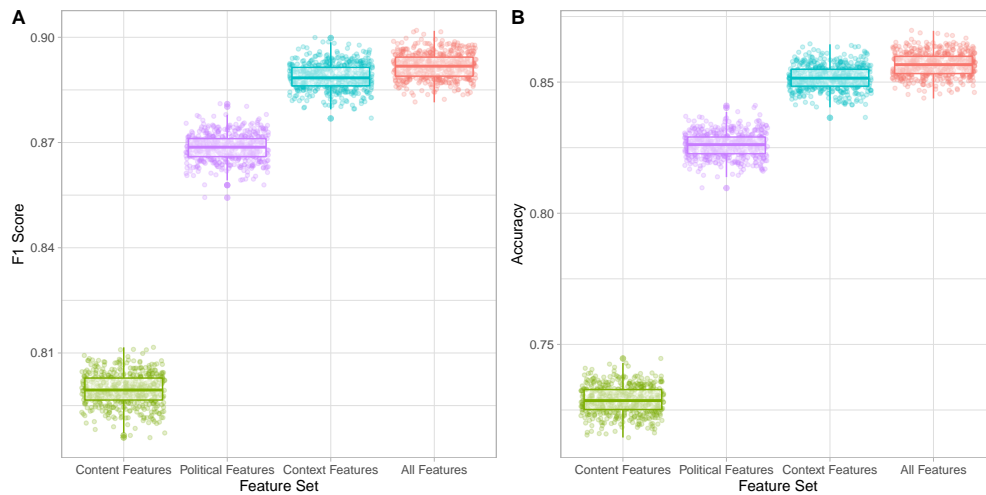


Figure A.2: A comparison of RF models predicting helpfulness classification of notes, trained with different sets of feature using (A) F1 Score and (B) Accuracy as Evaluation Metrics

Panel A shows a boxplot showing the performance of different models using F1 score as a metric, Panel B shows a boxplot showing the performance of different models using Accuracy as the evaluation metric

A.2 Logistic Regression, Full Results

Results for a model predicting misleadingness can be found in Table A.1; results for a model predicting helpfulness can be found in Table A.2.

Table A.1: Full logistic regression output for predicting misleadingness classifications of tweets. Table rendered via [169]

Constant	2.017*** (0.423)
Note writer Follower Count	-0.00000* (0.00000)
Note writer Statuses Count	0.00000 (0.00000)
Note writer Gender	0.319 (0.257)
Tweeter Follower Count	-0.00000* (0.000)
Tweeter Statuses Count	0.00000 (0.00000)
Tweeter Gender	0.151 (0.188)
Note writer Age	0.025 (0.089)
Tweeter Age	-0.059 (0.094)
Tweet Length	0.002 (0.001)
Tweet Sentiment	0.177 (0.144)
Tweet FK Score	-0.001 (0.002)
Tweet URL Count	-0.017 (0.131)
Note writer Partisanship Score	0.181 (0.126)
Tweeter Partisanship Score	-0.119 (0.120)
Note writer Partisanship Score X Tweeter Partisanship Score	-1.254*** (0.136)

Note:

*p<0.05; **p<0.01; ***p<0.001

Table A.2: Full logistic regression output for predicting helpfulness classifications of notes.

Constant	0.430 (0.294)
Note writer Follower Count	-0.000 (0.00000)
Note writer Statuses Count	0.00000 (0.00000)
Note writer Gender	-0.083 (0.102)
Tweeter Follower Count	-0.000** (0.000)
Tweeter Statuses Count	0.00000 (0.00000)
Tweeter Gender	0.016 (0.075)
Rater Follower Count	0.00000 (0.00000)
Rater Statuses Count	0.00000 (0.00000)
Rater Gender	-0.118 (0.164)
Note writer Age	-0.110* (0.047)
Tweeter Age	-0.038 (0.047)
Rater Age	-0.052 (0.040)
Tweet Length	-0.0004 (0.0005)
Note Length	0.002*** (0.0005)
Tweet Sentiment	0.058 (0.066)
Note Sentiment	0.182* (0.085)
Tweet FK Score	-0.001 (0.001)
Note SK Score	-0.00001 (0.0003)
Note URL Count	0.494*** (0.106)
Tweet URL Count	-0.048 (0.062)
Note writer Partisanship Score	-0.246** (0.075)
Rater Partisanship Score	0.206** (0.072)
Tweeter Partisanship Score	0.126* (0.050)
Note writer Partisanship Score X Rater Partisanship Score	1.268*** (0.074)
Note writer Partisanship Score X Tweeter Partisanship Score	-0.054 (0.045)
Rater Partisanship Score X Tweeter Partisanship Score	-0.517*** (0.058)
Note writer Partisanship Score X Rater Partisanship Score X Tweeter Partisanship Score	-0.073 (0.052)

Note:

*p<0.05; **p<0.01; ***p<0.001

Appendix B

Chapter 3 Appendix

B.0.1 Study Variable Definitions

Outcome Variable

The **Vaccine Intentions Index**, our main outcome for Studies 1 and 2, was composed as follows. All participants were asked the following 4 questions on a 0 to 100 scale, with 0 corresponding to “Definitely No” and 100 corresponding to “Definitely Yes”. We averaged the 4 questions together (where available) to create an index which was our main outcome of analysis.

- **Future Vaccine Intentions** All participants: “Imagine that a new COVID-19 strain, the Omega variant, arises. Imagine that Omega is able to evade the protection offered by current COVID-19 vaccines (or prior infection) - i.e., Omega achieves “immune escape.” In response, drug companies develop a new version of the COVID-19 vaccine that is effective against Omega. How likely would you be to get the new vaccine?”
- **Vaccine Intentions** If they had not received a COVID-19 vaccine already: “How likely are you to get the COVID-19 vaccine?”
- **Booster Intentions** If they had received a COVID-19 vaccine, but not a booster: “How likely are you to get a “booster” shot of a COVID-19 vaccine?”
- **Child Vaccine Intentions**
 - If they had a child (asked separately for children under 5 and children 5-18 due to differences in FDA approval). “Consider your child (under 5 / between 5 and 18). How likely is it that you would vaccinate your child with the COVID-19 vaccine?”
 - If they did not have a child: “Imagine that you had a child between 5 and 18 years old. How likely is it that you would vaccinate your child with the COVID-19 vaccine?”

Pre-Treatment Covariates

We fit our models using the following pre-treatment covariates in order to increase statistical power [170].

- *pre_vax_index* Our vaccination index defined above, measured pre-treatment
- *pol* Participant’s stated political leaning, measured on a 6-point scale from “Strong Republican” to “Strong Democrat”
- *gender* Participant’s stated gender, coded as 1 for female, else 0
- *age* Participant’s stated age measured on a continuous scale

B.0.2 Model

Study 1: Effect of Misinformation

To estimate the average treatment effect of vaccine misinformation on vaccination intentions, as well as the amount of heterogeneity between misinformation stimuli, we estimate the following multi-level model.

$$\begin{aligned}
 Y_i &= \beta_0 + \beta_1 \text{pre_vax_index}_i + \alpha \text{vax_treat}_i + \beta_2 \text{pre_vax_index}_i \times \text{vax_treat}_i \\
 &\quad + \beta_3 \text{gender}_i + \beta_4 \text{age}_i + \beta_5 \text{pol}_i + \epsilon_i \\
 \alpha &= \alpha_0 + \alpha_k \\
 \alpha_k &\sim N(0, \sigma)
 \end{aligned}$$

Where:

- i indexes subject
- k indexes stimulus
- Y_i is the post-vaccine index for subject i
- α_0 is the average treatment effect for vaccine misinformation
- α_k is the stimulus-level random effect for vaccine misinformation item k
- σ is the standard deviation of the stimulus-level random effects

Our quantities of interest are α_0 , the average treatment effect of misinformation, and σ , the degree of variation between misinformation items, measured as the standard deviation of the distribution of misinformation treatment effects.

Studies 1 and 2: Stimulus-Level Heterogeneity

To evaluate the extent to which content-level features predict variation in treatment effects, we perform the following two-stage process. Note that we specify the model with a single moderating variable x for readability, but the model could easily be specified as a vector of content-level features X as in a typical OLS regression model without loss of generality.

In Stage 1, we estimate the treatment effect $\hat{\theta}_{jk}$ for each stimulus k in study j by the following model. We estimate each study j separately; each subject i was assigned to only a single treatment k and a single experiment j .

For each study j , we estimate:

$$Y_{ij} = \beta_{0j} + \sum_{k \in K} \theta_{jk} \text{treat}_{ijk} + \beta_{1j} \text{pre_vax_index}_{ij} + \beta_{3j} \text{gender}_{ij} \\ + \beta_{4j} \text{age}_{ij} + \beta_{5j} \text{pol}_{ij} + \epsilon_{ij}$$

where:

- i indexes subject
- j indexes study
- k indexes stimulus
- treat_{ijk} is a dummy variable indicating whether individual i saw treatment k in study j
- Y_{ij} is the post-vaccine index for subject i in study j

In Stage 2, we run a meta-regression with our vector of estimated treatment effects $\hat{\theta}$ as our dependent variable and x , our content-level feature of interest, as our moderator. Because there is correlation between the treatment effects due to a common control group, we perform multi-variate meta-analysis and use the estimated variance-covariance matrix of $\hat{\theta}$, $\hat{\Sigma}$, from Stage 1 to parametrize η , the vector corresponding to sampling error at the subject level.

$$\hat{\theta}_{jk} = \beta_0 + \lambda x_{jk} + \xi_{(1)jk} + \xi_{(2)j} + \eta_{jk} \\ \xi_{(1)jk} \sim N(0, \sigma_1) \\ \xi_{(2)j} \sim N(0, \sigma_2) \\ \eta \sim N(0, \hat{\Sigma})$$

- j indexes study
- k indexes stimulus
- β_0 is intercept representing the baseline treatment effect
- x_{jk} is the stimulus-level characteristic of interest
- $\xi_{(1)jk}$ is the stimulus-level random effect
- $\xi_{(2)j}$ is the study-level random effect
- $\hat{\Sigma}$ is the block-diagonal variance-covariance matrix of $\hat{\theta}$ estimated in Part 1

Our quantity-of-interest is λ , the coefficient on our content-level feature x .

B.0.3 Balance Checks

To check for balance, for each pre-treatment covariate, we calculate the mean of the treatment and control group, respectively, and compare them using a t-test. Both studies show balance across covariates. We report both the unadjusted p -value and the adjusted p -value, after performing a Benjamini-Hochberg correction procedure. Experiment 1 is reported in Table B.1; Experiment 2 is reported in Table B.2.

Table B.1: Balance Check, Study 1

Variable	Control Mean	Treat Mean	p	p.adj
Is Female	0.56	0.55	0.47	0.73
Age	47.02	47.36	0.46	0.73
Pre Vaccine Index	64.34	65.14	0.41	0.73
Is Democrat	0.55	0.56	0.55	0.73
Is Unvaccinated	0.24	0.25	0.70	0.73
Is Boosted	0.54	0.54	0.73	0.73

Table B.2: Balance Check, Study 2

Variable	Control Mean	Treat Mean	p	p.adj
Is Female	0.51	0.52	0.35	0.49
Age	46.60	47.26	0.29	0.49
Pre Vaccine Index	60.10	61.05	0.49	0.49
Is Democrat	0.55	0.53	0.37	0.49
Is Unvaccinated	0.27	0.26	0.41	0.49
Is Boosted	0.51	0.53	0.31	0.49

B.0.4 Differential Attrition

Again, we test for differential attrition with a logistic regression model predicting whether or not a person attrited (1 = attrited, 0 = did not attrit) given 1) whether they were in the treatment vs. control group and 2) whether they were exposed to different features of treatment content.

Study 1 Attrition

The overall attrition rate is 2.9%. As can be seen in Table B.3, we find no evidence of differential rates of attrition in treatment vs. control. We also find no evidence of differential attrition by any features of the content; that is, people who were exposed to content of certain types (e.g. content that suggested the vaccine was harmful) were no more likely to drop out than people who saw content suggesting the vaccine was helpful to one’s health.

Study 2 Attrition

We test for differential attrition with a model predicting whether or not a person attrited (1 = attrited, 0 = did not attrit) given 1) whether they were in the treatment vs. control group and 2) whether they were exposed to different features of treatment content. The overall attrition rate is 1%.

As can be seen in Table B.4, we find no evidence of differential attrition by any features of the content. However, we do find evidence of attrition by whether or not the subject was

Table B.3: Attrition Check, Study 1

Variable	Coefficient	p.value	p.adj
Treatment (vs Control)	0.01	0.10	0.21
Harmful (vs Helpful) to Health	-0.01	0.18	0.21
Is Misinformation	0.01	0.10	0.21
Surprising	0.00	0.51	0.51
Pro Democrat (vs. Republican)	-0.01	0.03	0.21
Plausible	0.01	0.13	0.21
Familiar	-0.01	0.18	0.21

^a Note: For all content-level variables except treat, we filter to participants in the treatment condition only.

in treatment vs. control. People in the Control group were 2.9% more likely to attrit than in the treatment group (where attrition is very low – 0.7%). An examination found that this attrition was likely due to a technical error in the loading of content in the treatment group preventing members of the control group from advancing in the study. While this omission is unfortunate, analysis shows it is unlikely to affect our results. Analysis suggests this attrition is random; control-group attrition cannot be predicted from age, gender, vaccination status, political leaning, and pre-treatment vaccination intentions is non-significant ($F(920)=.87$, $p=.51$).

Furthermore, our quantity-of-interest in the second study is looking at differences in vaccine related treatments by features of the treatment content; the control group serves largely as a reference group. A Manski “worst-case” bound case analysis confirms this point. If we set the post-treatment vaccination intentions for all attriters in the treatment group to have an upper bound value of 100 and all attriters in the control group to have a lower bound value of 0, our meta-regression testing whether content that implies the vaccine is harmful to a person’s health is essentially unchanged and remains significant ($\beta=-.59, p=.007$).

Table B.4: Attrition Check, Study 2

Variable	Coefficient	p.value	p.adj
Treatment (vs Control)	-0.03	0.00	0.00
Harmful (vs Helpful) to Health	0.00	0.30	0.64
Is Misinformation	0.00	0.43	0.64
Surprising	0.00	0.37	0.64
Pro Democrat (vs. Republican)	0.00	0.60	0.65
Plausible	0.00	0.65	0.65

^a Note: For all content-level variables except treat, we filter to participants in the treatment condition only.

B.0.5 Additional Survey Results

Post-hoc, we collected additional labels from Lucid to assess whether dimensions of content that we missed predicted variation in the effects. These additional moderators are shown in Figure B.1, along with the logged engagement (`engagementL`) that the URL received on Facebook, collected from CrowdTangle. The only additional variables that explained the effect were harm-related, providing additional support for our main results.



Figure B.1: The coefficients from respective meta-regressions testing how different features of content moderate the treatment effects in Study 2. Coefficients with $p < .05$ are bolded.

B.0.6 Subject Level Heterogeneity

In Study 1, we examine subject-level heterogeneity in the treatment effect by separately testing for an interaction between treatment (misinformation vs. control) and a given subject-level characteristic. The results are shown in Table B.5.

We find no strong evidence of individual-level heterogeneity. Our pre-registered model included an interaction term between treatment and pre-treatment vaccination intentions which was not significant ($\beta = -.34$, $p = .31$).

Post-hoc, we also examined heterogeneity by gender, age, political leaning, vaccination status, and booster status. Although both gender and unvaccinated status suggest a possible difference ($p = .07$), after adjusting for multiple comparisons, the evidence for substantial heterogeneity is unconvincing. Furthermore, an omnibus test in the following section also fails to reject the null of subject-level heterogeneity, see Section B.0.6. This low degree of

subject-level heterogeneity is consistent with past political science research that finds low amounts of individual-level heterogeneity with regards to political persuasion [153].

Table B.5: Study 1: Individual Level Heterogeneity

study	Variable	Estimate	p.value	p.adj
Study 1	Is Male	0.70	0.69	0.74
Study 1	Age (Standardized)	0.29	0.74	0.74
Study 1	Is Democrat (vs. Republican)	-1.34	0.45	0.74
Study 1	Is Unvaccinated	3.04	0.06	0.36
Study 1	Pre-Treatment Vaccination Intentions (Standardized)	-0.34	0.31	0.74
Study 1	Is Boosted	0.66	0.65	0.74

We repeat the same post-hoc heterogeneity analysis for Study 2. The results are shown in Table B.6. Note that in Study 2, participants were randomized to see either misinformation vaccine content, true vaccine content, or control content. Thus, we coded treatment as "1" if the participant was randomized to see misinformation content, and 0 if the participant was randomized to see control or true content.

Similarly, we find no strong evidence for individual-level heterogeneity in susceptibility to misinformation in Study 2.

Table B.6: Study 2: Individual Level Heterogeneity

study	Variable	Estimate	p.value	p.adj
Study 2	Is Male	2.23	0.40	0.71
Study 2	Age (Standardized)	-0.23	0.86	0.99
Study 2	Is Democrat (vs. Republican)	-4.31	0.11	0.57
Study 2	Is Unvaccinated	3.18	0.19	0.57
Study 2	Pre-Treatment Vaccination Intentions (Standardized)	-0.58	0.24	0.57
Study 2	Is Boosted	-0.53	0.81	0.99
Study 2	NA	0.04	0.99	0.99

Subject Level Heterogeneity, Causal Forest

Post hoc, we also run a causal forest model [171] with Pre-Treatment Vaccination Intentions, Age, Gender, and Political Lean as predictors. As treatment variables, we separately consider 1) whether the treatment was misinformation and 2) a continuous variable for the "harmful-to-health" rating of the treatment. Because Control treatment were not originally labeled, we give them the rating "3", corresponding to "Neither harmful nor helpful". We run the causal forests for Experiment 1 and 2, respectively.

As shown in Table B.7, an omnibus test finds no significant heterogeneity across any of the specifications [172]. The predicted treatment effects from the causal forest, shown in Figure B.2, also show little evidence of heterogeneity.

Table B.7: P-values from omnibus test of heterogeneity from causal forest models

Experiment	Treatment Var	Omnibus-Heterogeneity-Test p-value
Experiment 1	Is Misinformation	0.49
Experiment 2	Is Misinformation	1.00
Experiment 1	Harmful to Health	0.78
Experiment 2	Harmful to Health	1.00

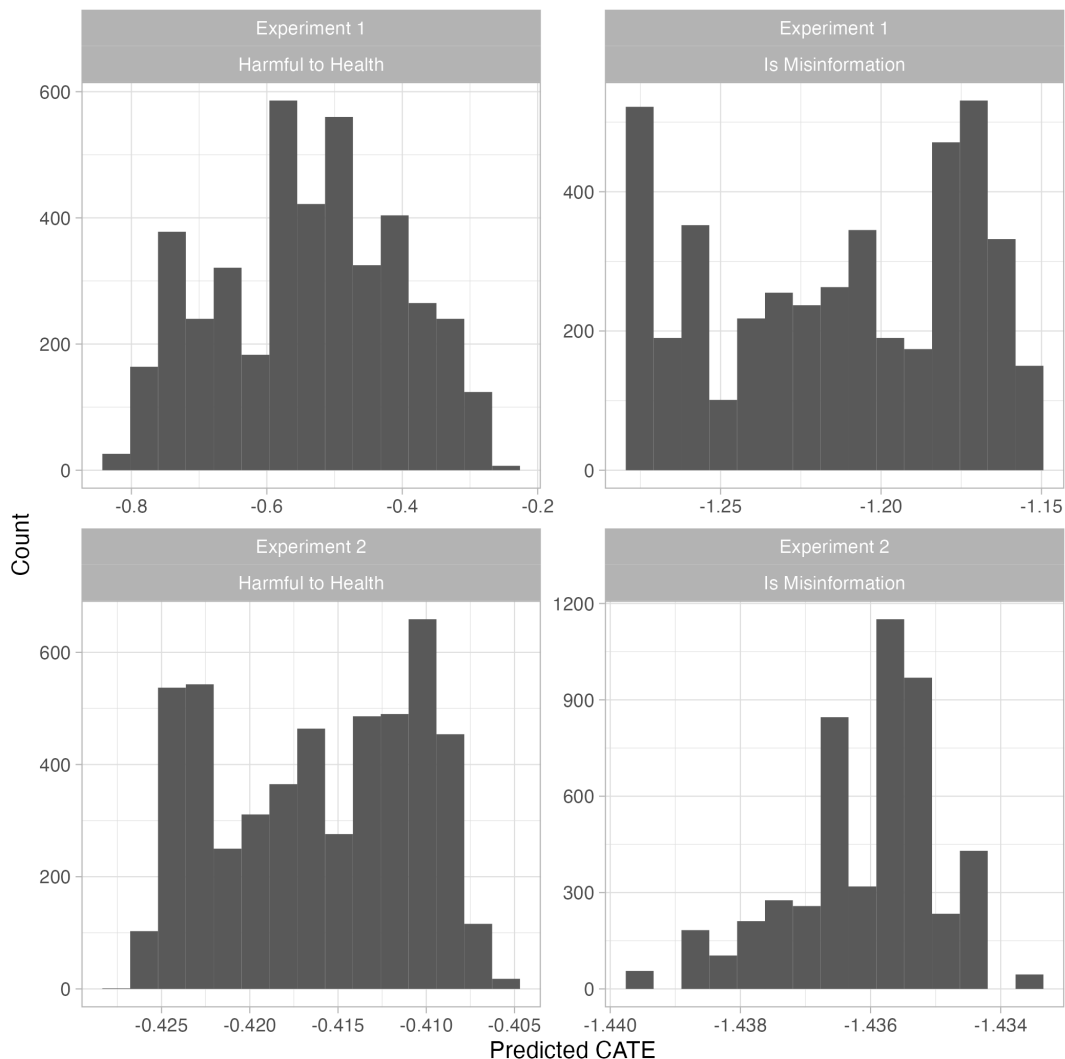


Figure B.2: The predicted CATEs from causal forest model for different experiments and treatments, respectively.

B.0.7 Cutoff Tuning

For each URLs for which we have ground-truth labels, we classify a URL “vaccine-skeptical” if it has a ground-truth “Crowdsourced Aggregate Score” less than 3, the scale midpoint, otherwise we classify it as “not vaccine-skeptical”. We use the term “hesitancy-inducing”

interchangeably with “vaccine-skeptical.”

Then, to assess the performance of our model on our binary classification task, we use the predicted “Crowdsourced Aggregate Score” to predict the binary “vaccine-skeptical vs. not skeptical / promotional” class of each URL. Because we are using the continuous “Crowdsourced Aggregate Score” to predict a binary classification, we have to pick a threshold of the score below which we predict the URL is “vaccine-skeptical.” Figure B.3 shows how the false-positive-rate, true-positive-rate, and accuracy of the model varies at different cutoffs of the predicted “Crowdsourced Aggregate Score.” Based on these performance metrics, we choose “3” as a cutoff for the predicted “Crowdsourced Aggregate Score,” which also has the benefit of being the same threshold we used for our ‘ground truth data. A cutoff of “3” has a high accuracy (91%), a low false-positive-rate (4%), and a high-true-positive-rate (84%). We could have chosen a model with a slightly higher accuracy and true-positive-rate, but we chose a cutoff model with a low false-positive-rate to guard against URLs that are “Not-vaccine-skeptical / promotinhg” (and potentially even promoting vaccination) being considered “vaccine-skeptical” (i.e. questioning vaccination).

B.0.8 Top Viewed URLs

We show in Figure B.4 the results top individual URLs, rather than the top story clusters. The results are similar to the top clusters (e.g. the “Healthy Doctor Died...” from the Chicago Tribune is also the most viewed URL). One noticeable difference is that there are 5 stories from Unicef.org in the top URLs. These URLs are markedly different from the other stories, which covered news events or important safety information. These URLs were either part of the second-largest cluster, which included information about Covid safety, or part of the cluster corresponding to niche or tangential stories which we excluded from the main analysis.

These Unicef stories received substantially less engagement-per-view than other top stories (0.2% engagement per view, compared to 4.5% for other top stories – a 20 fold difference). We suspect that these stories were likely shown to viewers as part of the “Covid Information Hub,” a product by Facebook that was pinned to top of newsfeed and featured information from Unicef and other nonprofit organizations. We show the top 10 URLs with and without these stories.

B.0.9 Most Harmful Domains

In Figure B.5, we rank the top most harmful domains by overall impact. We calculate the total overall impact of each domain by the following process. First, we subset to URLs predicted to be vaccine-sketptical (i.e. with a “Crowdsourced Aggregate Score” less than 3, the scale midpoint). Then, for each URL, we compute the total impact as the number of views times the predicted persuasive impact, conditional on viewership. Finally, we sum overall URLs for each domain and normalize by the total number of US Facebook viewers. Note that this ranking is only based on the predicted negative impact from vaccine-skeptical stories; we do not consider the potential positive impact from stories promoting the vaccine because promotional vaccine stories did not increase vaccination intentions in our survey experiments.

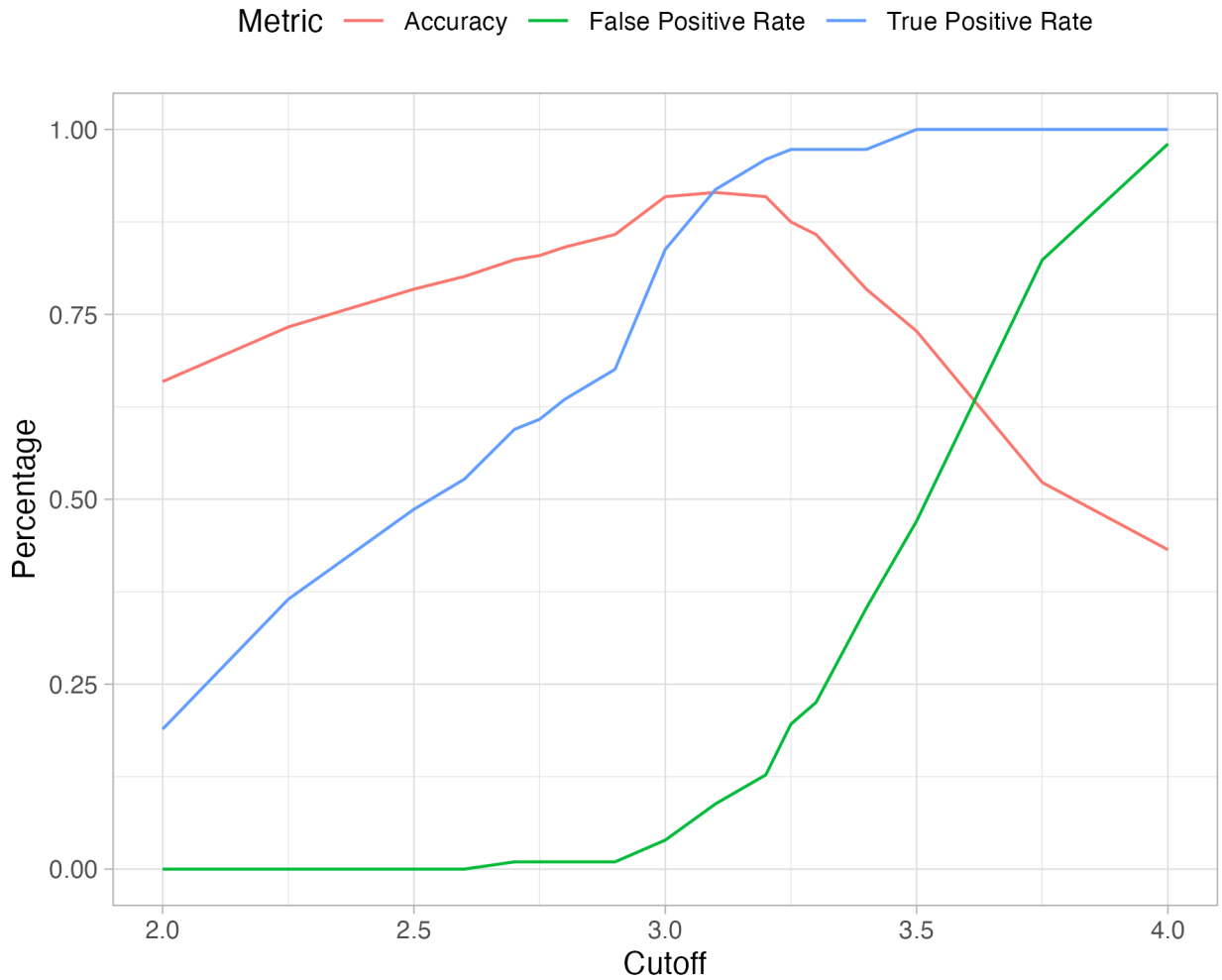


Figure B.3: Performance of a model that uses the continuous predicted “Crowdsourced Aggregate Score,” binarized at varying thresholds, to predict the binary “Vaccine-skeptical” vs “Not skeptical / promoting” rating of a URL. We use the term “hesitancy-inducing” interchangeably with vaccine-skeptical.

As can be seen, the most harmful domains are all popular mainstream domains, like the *Chicago Tribune* or *The New York Post*. Even *The New York Times* had a substantial negative impact. An inspection shows that these high-quality domains had significant reach and devoted coverage to rare vaccine deaths and side effects. For example, *The New York Times* published two stories on the Miami doctor with the headlines “The death of a Miami doctor who received a coronavirus vaccine is being investigated” and “Doctor’s Death After Covid Vaccine Is Being Investigated” that received widespread views.

In Figure B.6, we examine the top most harmful (i.e. hesitancy-inducing) domains, ranked by the predicted persuasive effect of the average URL. Unlike Figure B.5, this ranking does not weight impact by the number of views each URL received. Panel A shows the ranking over all domains, and Panel B shows the ranking for all domains that published at

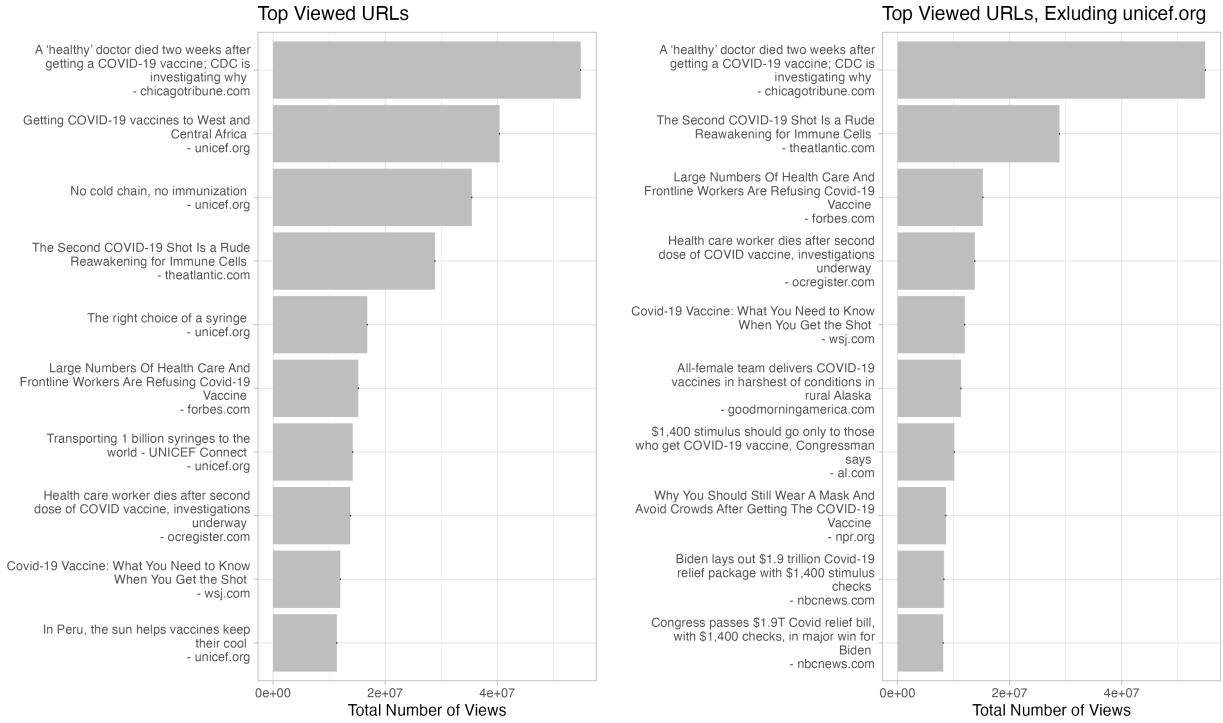


Figure B.4: Top Individual URLs. Panel A shows the Top URLs including unicef.org. Panel B shows the Top URLs excluding unicef.org.

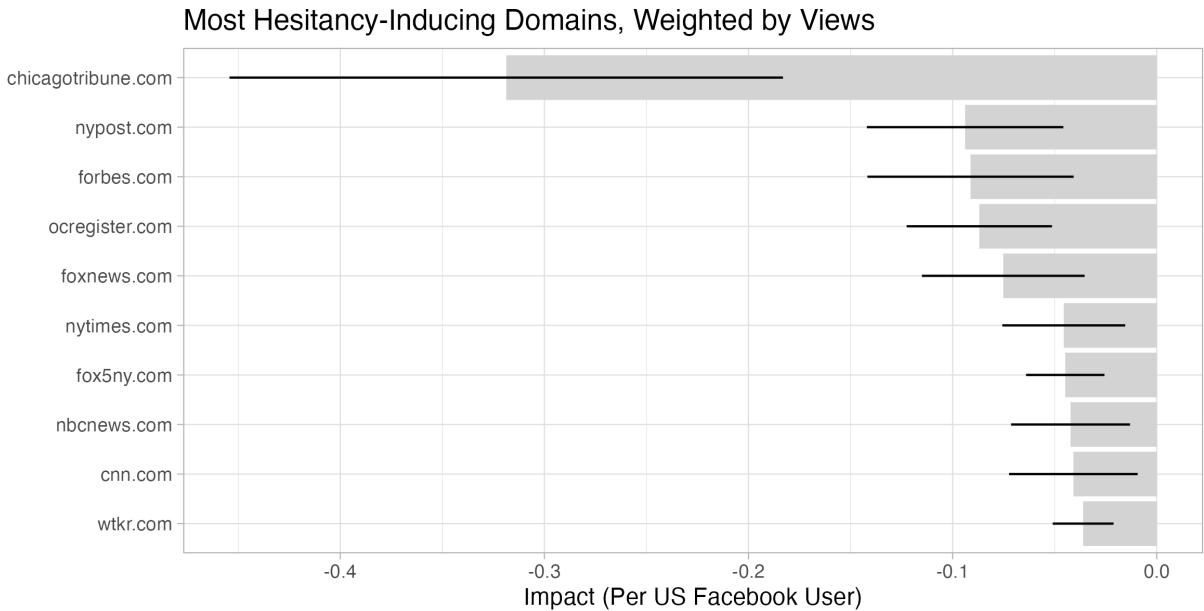


Figure B.5: Top Harmful Domains, Weighted by Views

least 20 URLs in our full set of 13,206.

Unlike Figure B.5, the ranking are dominated by little-known fringe sources or low-credibility fake news domains, such as infowars.com and childrenshealthdefense.org, a site

run by the anti-vaccine politician Robert F. Kennedy Jr. These sites received much less viewership than the most popular mainstream domains; however, conditional on viewership, their content was much more negatively impactful.

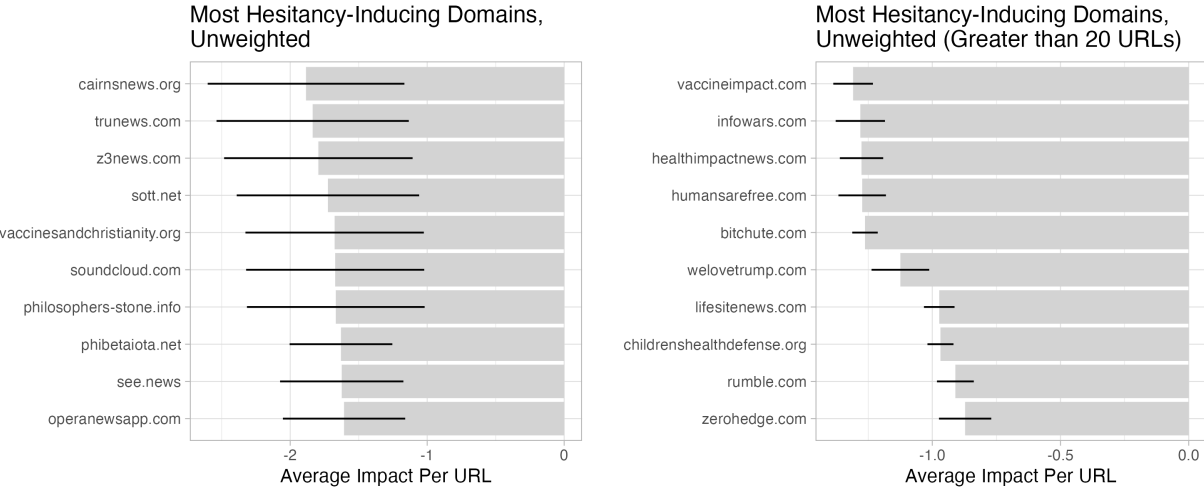


Figure B.6: Top harmful (i.e. hesitancy-inducing) domains, based on average URL treatment effect. **Left** The top most harmful domains across all domains. **Right** The top most harmful domains among domains that have at least 20 URLs in the dataset

B.0.10 Contemporaneous Treatment Effect Estimates

As described in Section B.0.15, we calculate our total impact estimates for URLs that we classify as “vaccine-skeptical” based on whether they have a “Crowdsourced Aggregate Score” s_i less than 3. Content with $s_i \geq 3$ are excluded from analysis.

In our original analysis, we find little evidence that content that promoted vaccination (i.e. with $s_i \geq 3$) actually increased vaccination intentions. Yet, it is possible that this null effect is due to the fact that by the time that we ran our experiments (in mid-2022), exposure to pro-vaccine content was already saturated and people had formed strong prior opinions about their willingness to take a vaccine, such that an additional marginal exposure to pro-vaccine content had no detectable effect. Such an account is consistent with Bayesian explanations of persuasion, which have shown that people show larger magnitude changes in opinion on topics on which they have less prior knowledge [173]–[175]. In early 2021, during the rollout of the vaccine, it is likely that the environment was less saturated with pro-vaccine content, and thus, pro-vaccine content might have been more persuasive.

Therefore, we consider how our estimate of the overall impact of vaccine content on vaccination intentions would change given alternative estimates for promotional vaccine content. Figure B.7 estimates the net impact for promoting and vaccine-skeptical content at various cutoffs for whether or not a URL is considered “Vaccine-skeptical” vs. “Not-Skeptical / Promoting”, and for varying estimates of the effect for a single exposure to an item of vaccine-promoting content.

As an example, Athey and colleagues [117] tested the impact of pro-vaccine ads on willingness to get a vaccine on Facebook in early 2021 and found that this promotional content

had a statistically insignificant effect on vaccination intentions of 0.1 percentage points. This corresponds to the yellow line in Figure B.7. If we use a skeptical threshold of 3 as in the main text and assume that promoting vaccine content has an positive average impact of 0.1pp per exposure, then we estimate that pro-vaccine content increased vaccination intentions by .8pp per Facebook user. This suggests that the overall net impact of vaccine content on Facebook would be -1.5pp per user, as opposed to -2.3pp.

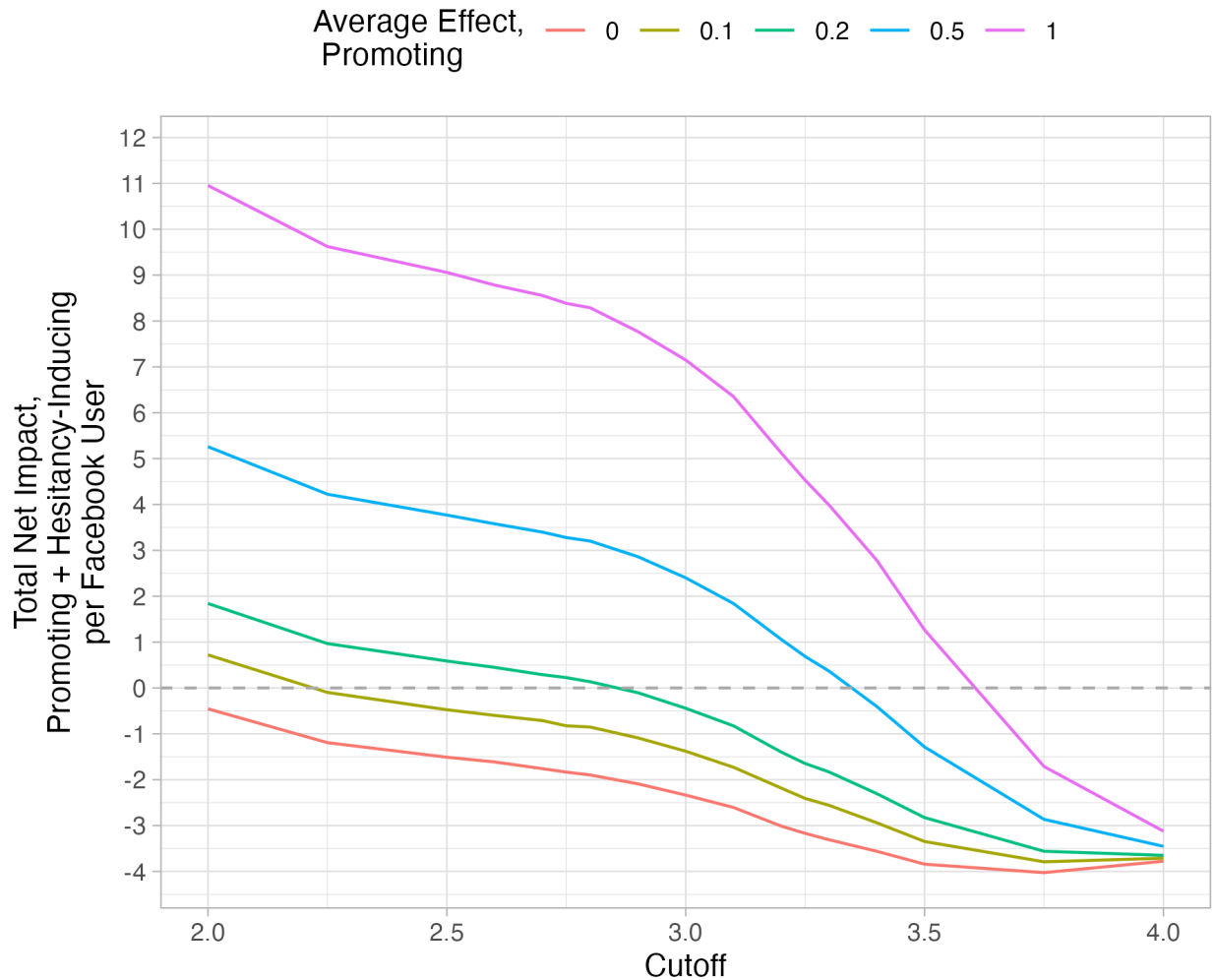


Figure B.7: Net impact of promoting and vaccine-skeptical vaccine content as a function of the cutoff of the Crowdsourced Aggregate Score. “Cutoff” is the threshold of the score at which a URL is classified as either 1) vaccine-skeptical or 2) promoting. The colored lines show the net impact at different values for the average promoting effect. “Hesitancy-inducing” is used interchangeably with “vaccine-skeptical.”

At the same time, it is also possible that anti-vaccine misinformation and vaccine-skeptical content might have had also had larger negative effect on vaccination intentions in early 2021 than in mid 2022. In a study run in September 2020, before the rollout of the vaccines, Loomba et al (2021) found that anti vaccine misinformation lowered intentions to take a COVID-19 vaccine by 6.4 percentage points [128]. This is 4.25X the size of the 1.5 percentage points average effect of misinformation on vaccination intentions we find in our experiments – although we note that each participant saw 5 pieces of misinformation in that experiment, making the estimates closer than they initially appear. Nonetheless, scaling the magnitude of our results by 4.25 would suggest that vaccine-skeptical misinformation and non-misinformation content lowered overall vaccination intentions on Facebook by 9.9 percentage points, rather than the 2.33 percentage points that we estimated. Combining that with the +.8pp increase in vaccination intentions from pro-vaccination content, we would estimate that cumulatively, content on Facebook lowered vaccination intentions by 9.1pp.

Different assumptions about the persuasive effect of promoting and skeptical vaccine content in 2021 would naturally yield different results. Here, we present a range of possible results under reasonable sets of assumptions; however, we hope that future research might extend our framework to different assumptions.

B.0.11 Formal Model

We use a two-step model to evaluate how well crowdsourced judgments predict our vaccine treatment effects. This is the same model as in Section B.0.2, but s refers to the “Crowdsourced Aggregate Score”, rather than any content-level feature.

We estimate our 130 treatment effects $\hat{\theta}_{jk}$ using the same fixed-effect regression with HC2 robust standard errors, as defined in Section B.0.2.

Then, for our set of $\hat{\theta}_{jk}$, we estimate the following model using a random effects meta-regression:

$$\begin{aligned}\hat{\theta}_{jk} &= \beta_0 + \beta_1 s_{jk} + \xi_{(1)jk} + \xi_{(2)j} + \eta_{jk} \\ \xi_{(1)jk} &\sim N(0, \sigma_1) \\ \xi_{(2)j} &\sim N(0, \sigma_2) \\ \eta &\sim N(0, \hat{\Sigma})\end{aligned}$$

- j indexes study
- k indexes stimulus
- β_0 is intercept representing the baseline treatment effect
- s_{jk} “Crowdsourced Aggregate Score” for stimulus i in study j
- $\xi_{(1)jk}$ is the stimulus-level random effect
- $\xi_{(2)j}$ is the study-level random effect
- $\hat{\Sigma}$ is the block-diagonal variance-covariance matrix of $\hat{\theta}$ estimated in Part 1

Our quantity of interest is the coefficient on our “Crowdsourced Aggregate Score” β_1 and I^2 of the model.

B.0.12 Alternative Models

We also report the results for two alternate models in Table B.8 and Table B.9. "Single" is a model in which we trained a single model to predict the "Crowdsourced Aggregate Score" directly, instead of training 3 models to predict each component separately and then averaging them together post-hoc (a "composite" model). "Clustered" is a model in which we trained a composite model using a training procedure in which we first clustered our input headlines and descriptions in the CT-BERT embedding space, and then held out clusters, rather than individual URLs, to guard against data leakage. "Composite" is the best-performing model that predicts each component of our "Crowdsourced Aggregate Score" separately, and then averages them together in an ensemble-style method. "Single" and "Composite" both use the same train/test split: a random 85/15 the URL level, stratified by the "Less vs. More Vax" value.

As can be seen, all models / training procedures have similar performance. Because the "Composite" model has slightly better performance and is more conservative (i.e. has a lower false-positive rate), we choose it for our model in our main text.

Table B.8: Performance Metrics for Alternate Methods

Variable	MSE	RMSE	MAE	Accuracy (with .5)	Accuracy (with 1)
Crowdsourced Aggregate Score (Clustered)	0.13	0.36	0.27	0.86	0.99
Crowdsourced Aggregate Score (Composite)	0.11	0.34	0.26	0.86	0.99
Crowdsourced Ag- gregate Score (Sin- gle)	0.12	0.35	0.27	0.86	0.99

B.0.13 Impact Calculation

We calculate the total impact-per-user of vaccine-skeptical i) vaccine-skeptical and ii) misinformation content, respectively, based on the following toy model.

For each vaccine-related Facebook URL i , we have the following variables:

- p_i : The predicted persuasive effect for each URL i .
- v_i : The number of views for each URL i .
- s_i : The "Crowdsourced Aggregate Score" for each URL i .

Table B.9: Performance Metrics for Alternate Methods, Binary Classification Task

Variable	Accuracy	AUC	F1-Score	FPR	TPR
Crowdsourced Aggregate Score (Clustered)	0.88	0.95	0.86	0.11	0.86
Crowdsourced Aggregate Score (Composite)	0.91	0.97	0.89	0.04	0.84
Crowdsourced Aggregate Score (Single)	0.86	0.96	0.82	0.05	0.74

- h_i : Binary indicator for whether i is vaccine-skeptical (1) or Not (0). We classify a URL as “vaccine-skeptical” if $s_i < 3$, the scale midpoint (for an exploration of other cutoffs see Figure B.3).
- m_i : Binary indicator for whether URL i is fact-checked as misinformation (1) or not (0).

We then define two sets of URLs U_{VS} and U_M :

- U_M : The set of URLs designated as vaccine-skeptical and contain misinformation.
- U_{VS} : The set of all URLs designated as vaccine-skeptical but do not contain misinformation.

Formally, this is:

$$U_M = \{i | h_i = 1 \text{ and } m_i = 1\}$$

$$U_{VS} = \{i | h_i = 1 \text{ and } m_i = 0\}$$

Given the total number of 2021 US Facebook users N_{FB} , we can calculate the the total impact per user for vaccine-skeptical misinformation URLs, I_M , and for vaccine-skeptical non-misinformation URLs I_{VS} as follows:

$$I_M = \frac{\sum_{i \in U_M} p_i \cdot v_i}{N_{FB}}$$

$$I_{VS} = \frac{\sum_{i \in U_{VS}} p_i \cdot v_i}{N_{FB}}$$

These equations represent the sum of the product of the predicted persuasive effect and the number of views for each vaccine-skeptical URL, divided by the total number of Facebook users, calculated separately for URLs that contain outright misinformation and for URLs that are vaccine-skeptical but do not contain outright misinformation.

B.0.14 Confidence Intervals for Impact Estimates

For each draw j from 1...1000 of our predicted treatment effects $p_{i,j}^*$ defined in Section ??, we calculate overall impact I_M^* and I_{VS}^* .

From $j = 1...1000$:

$$I_{M,j}^* = \frac{\sum_{u \in U_M} p_{i,j}^* \cdot v_i}{N_{FB}}$$
$$I_{VS,j}^* = \frac{\sum_{i \in U_{VS}} p_{i,j}^* \cdot v_i}{N_{FB}}$$

We report the full distributions as well as the 95% confidence intervals for I_M^* and I_{VS}^* , respectively, in Figure 4 of the main text.

B.0.15 Threshold for Skeptical URLs

For our impact estimates, we subset to URLs classified as skeptical, where we classify URL i as skeptical if the ‘‘Crowdsourced Aggregate Score’’ s_i is less than 3. In Figure B.8, we show how results would differ for different cutoffs for skeptical. The left panel shows the number of URLs designated as Skeptical vs. Neutral/Promoting of Vaccines at different cutoffs of the Crowdsourced Aggregate Score s . The right panel shows how our overall impact estimates (the sum of the impact of vaccine-skeptical misinformation I_M , and non-misinformation I_{VS} , defined in Section B.0.13), change as the cutoff for being classified as skeptical increases. As our cutoff increases, the number of URLs classified as skeptical increases. However, as our impact estimates increase with higher thresholds (since more URLs are included and considered skeptical), the width of our confidence intervals increase at higher thresholds as well, since a wider range of URLs – with a wider range of predicted effects, some of which have a positive upper bound – are included in our analysis.

B.0.16 Facebook Subject-Level Heterogeneity

In Figure B.9, we examine how vaccine coverage differed among different demographic groups on Facebook. In particular, we examine how exposure to vaccine-related content differs by gender, age bracket, and political-leaning – the three demographics made available to researchers via Facebook URL Shares dataset. Within each demographic bucket, we calculate the total percentage of URL views going to ‘‘skeptical’’ vaccine content compared to ‘‘Not skeptical / Promoting’’ content, where we classify a URL as ‘‘skeptical’’ if it has a ‘‘Crowdsourced Aggregate Score’’ less than 3. Skeptical content includes both vaccine-skeptical mainstream / accurate content as well as anti-vaccine misinformation. We report the proportion of vaccine content that is vaccine-skeptical, rather than the impact-per-user of vaccine-skeptical content on vaccination intentions, because Facebook does not publish the total number of users in each demographic bucket.

These results show that as one might expect, users who are 1) younger and 2) more conservative see relatively more vaccine-skeptical vaccine content than users who are older or more liberal. We find little evidence of differences between genders. Surprisingly and perhaps most concerningly, users who are non-political (i.e. who do not have a ‘‘Political

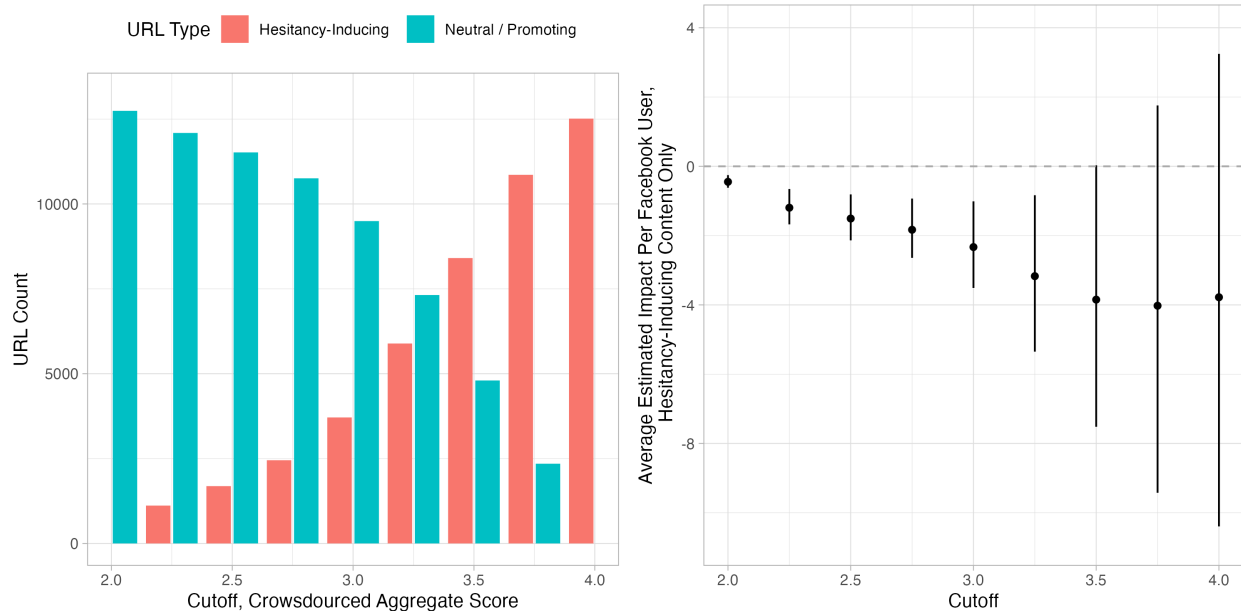


Figure B.8: Varying cutoffs impact on results. We use the term “hesitancy-inducing” interchangeably with “vaccine-skeptical.” **Left** How varying the cutoff affects the number of URLs considered vaccine-skeptical. **Right** How varying the cutoff affects overall impact estimates on vaccination intentions.

Page Affinity score”) are exposed to a substantial amount of vaccine-skeptical content. These findings suggest that exposure to vaccine-skeptical content is relatively common and not concentrated among certain demographics.

B.0.17 Crowdsourcing Variable Definition

Our “Crowdsourced Aggregate Score” was composed of the average of “Less vs. More Willing to Vaccinate”, “Harmful to Health”, and “Accuracy.” These three variables are described below.

- **Less vs. More Willing to Vaccinate** Do you think the above headline would make people less likely, or more likely, to take a vaccine for Covid-19? 1 - Much Less Likely to 5 - Much More Likely)
- **Harmful to Health (Reversed)** Does the above headline suggest the Covid-19 vaccine could be harmful or helpful to a person’s health? (1 - Very Helpful to 5 - Very Harmful). We reverse this score such 5 = Very Helpful, and 1 = Very Harmful, for consistency with the other 2 variables.
- **Accuracy** Adapted from [allen_scaling_2021](#). An average of the following 7 questions, rescaled to a 1-5 range for consistency with the other two variables.
 - Do you think this headline is accurate? (1 - Definitely No to 7 - Definitely Yes)
 - Do you think this headline is objective? (1 - Definitely No to 7 - Definitely Yes)

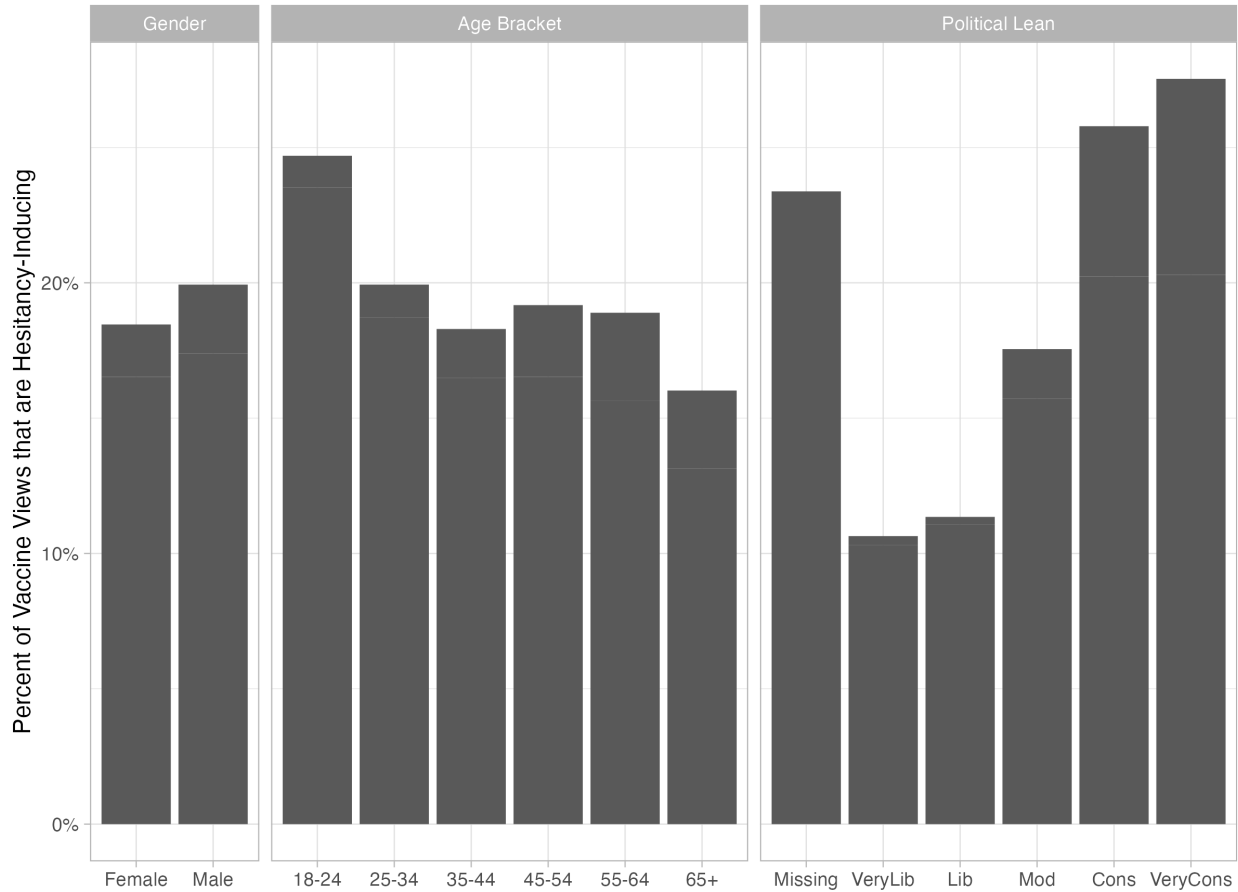


Figure B.9: Percent of total vaccine views going to “vaccine-skeptical” URLs for various demographics groups. URLs are classified as “vaccine-skeptical” if they have a “Crowdsourced Aggregate Score” less than 3.

- Do you think this headline was written in an unbiased way? (1 - Totally Biased to 7 - Totally Unbiased)
- Do you think this story describes an event that actually happened? (1 - Definitely No to 7 - Definitely Yes)
- Do you think this story is reliable? (1 - Definitely No to 7 - Definitely Yes)
- Do you think this story is trustworthy? (1 - Definitely No to 7 - Definitely Yes)
- Do you think this story is true? (1 - Definitely No to 7 - Definitely Yes)

B.0.18 Crowdsourcing Performance

In Table B.10, we compare the i) the baseline random effects meta-regression model (i.e. with no moderator), ii) the model with a single crowdsourced variable as a moderator (Less vs More Willing to Vaccinate, defined in Section B.0.17), iii) the model with three separate crowdsourced variables as moderators (Less vs. More Willing to Vaccinate, Harmful-to-

health, and Accuracy) and iv) the model with an aggregate crowdsourced variable as a moderator (the "Crowdsourced Aggregate Score", see above).

For meta-regressions, the pseudo- R^2 is defined as the proportional reduction in τ^2 between the baseline random effects model and the mixed effects model (i.e. with moderators), where τ^2 is the residual variance attributable not attributable to sampling error [176]. For our analysis, we define pseudo- R^2 in the proportional reduction in σ_1^2 , the variance attributed to the stimulus-level random effect.

$$R^{2*} = \frac{\tau_{RE}^2 - \tau_{ME}^2}{\tau_{RE}^2}$$

The crowdsourcing models show large improvement over the baseline random effects model ($\Delta\text{AIC} = 11.08, 14.08, \text{ and } 13.65$, respectively, well above the cutoff of 2 established in [177]). The multiple-question models also shows better fit than the single question model ($\Delta\text{AIC} = 2.61 \text{ and } 2.18$, respectively). In our main analysis, we choose the multiple question model with the three variables averaged into the "Crowdsourced Aggregate Score", rather than included separately, because the "Crowdsourced Aggregate Score" model is more parsimonious and has substantially higher I^2 and pseudo R^2 values. While the aggregate variable model has slightly higher AIC than the separate variable model, the difference is negligible.

	Baseline	Single Question	Multiple Questions, Sep	Multiple Questions, Avg
Intercept	-0.82 (0.55)	-2.67*** (0.62)	-0.53 (2.36)	-3.13*** (0.64)
Less vs. More Likely to Vaccinate		0.67*** (0.19)	0.20 (0.38)	
Harmful to Health			-0.42 (0.37)	
Accuracy			0.16 (0.21)	
Crowdsourced Aggregate Score				0.85*** (0.22)
Pseudo-R ²	0	0.31	0.26	0.34
I ²	37.6%	18.51%	15.09%	13.44%
τ^2	0.72	0.27	0.21	0.18
DF Resid.	129	128	126	128
AIC	430.39	418.92	416.31	416.74

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Table B.10: Model Comparison, Crowdsourcing

References

- [1] Facebook, *Facebook’s Third-Party Fact-Checking Program*. URL: <https://www.facebook.com/journalismproject/programs/third-party-fact-checking>.
- [2] Twitter, *Updating our approach to misleading information*. URL: https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html.
- [3] *How Google and YouTube are investing in fact-checking*, en-us, Nov. 2022. URL: <https://blog.google/outreach-initiatives/google-news-initiative/how-google-and-youtube-are-investing-in-fact-checking/> (visited on 05/10/2023).
- [4] T. Wood and E. Porter, “The elusive backfire effect: Mass attitudes’ steadfast factual adherence,” en, *Polit. Behav.*, vol. 41, no. 1, pp. 135–163, Mar. 2019.
- [5] 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,” *Manage. Sci.*, Feb. 2020.
- [6] M. Mali, *Critics fear Facebook fact-checkers losing misinformation fight*. The Hill, Jan. 2020. URL: <https://thehill.com/policy/technology/478896-critics-fear-facebook-fact-checkers-losing-misinformation-fight>.
- [7] D. Flamini, *Most Republicans don’t trust fact-checkers, and most Americans don’t trust the media - Poynter*. Jul. 2019. URL: <https://www.poynter.org/ifcn/2019/most-republicans-dont-trust-fact-checkers-and-most-americans-dont-trust-the-media/>.
- [8] G. Pennycook and D. G. Rand, “Fighting misinformation on social media using crowdsourced judgments of news source quality,” en, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 116, no. 7, pp. 2521–2526, Feb. 2019.
- [9] 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, Publisher: Sage Publications Sage CA: Los Angeles, CA.
- [10] J. Allen, B. Howland, M. Mobius, D. Rothschild, and D. J. Watts, “Evaluating the fake news problem at the scale of the information ecosystem,” en, *Sci Adv*, vol. 6, no. 14, eaay3539, Apr. 2020.
- [11] A. Guess, J. Nagler, and J. Tucker, “Less than you think: Prevalence and predictors of fake news dissemination on Facebook,” en, *Sci Adv*, vol. 5, no. 1, eaau4586, Jan. 2019.

- [12] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer, “Fake news on Twitter during the 2016 U.S. presidential election,” en, *Science*, vol. 363, no. 6425, pp. 374–378, Jan. 2019.
- [13] N. Dias, G. Pennycook, and D. G. Rand, “Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media,” *Harvard Kennedy School Misinformation Review*, vol. 1, no. 1, 2020.
- [14] 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,” en, *Cognition*, vol. 188, pp. 39–50, Jul. 2019.
- [15] G. Pennycook and D. G. Rand, “Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking,” en, *J. Pers.*, vol. 88, no. 2, pp. 185–200, Apr. 2020.
- [16] P. Moravec, R. Minas, and A. R. Dennis, *Fake News on Social Media: People Believe What They Want to Believe When it Makes No Sense at All*, en, SSRN Scholarly Paper, Rochester, NY, Aug. 2018. DOI: [10.2139/ssrn.3269541](https://doi.org/10.2139/ssrn.3269541). URL: <https://papers.ssrn.com/abstract=3269541> (visited on 04/04/2024).
- [17] M. W. Kattan, C. O’Rourke, C. Yu, and K. Chagin, “The Wisdom of Crowds of Doctors: Their Average Predictions Outperform Their Individual Ones,” en, *Med. Decis. Making*, vol. 36, no. 4, pp. 536–540, May 2016.
- [18] J. Surowiecki, *The Wisdom Of Crowds*, en. Anchor Books, 2005.
- [19] F. Galton, “Vox Populi,” en, *Nature*, vol. 75, no. 1949, pp. 450–451, Mar. 1907.
- [20] M. Gabielkov, A. Ramachandran, A. Chaintreau, and A. Legout, “Social Clicks: What and Who Gets Read on Twitter?” In *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science*, ser. SIGMETRICS ’16, New York, NY, USA: Association for Computing Machinery, Jun. 2016, pp. 179–192.
- [21] S. M. Herzog and R. Hertwig, “The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping,” en, *Psychol. Sci.*, vol. 20, no. 2, pp. 231–237, Feb. 2009.
- [22] S. Frederick, “Cognitive Reflection and Decision Making,” *J. Econ. Perspect.*, vol. 19, no. 4, pp. 25–42, Dec. 2005.
- [23] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. 2013.
- [24] C. Lim, “Checking how fact-checkers check,” *Research & Politics*, vol. 5, no. 3, 2018.
- [25] U. K. H. Ecker, S. Lewandowsky, and M. Chadwick, “Can Corrections Spread Misinformation to New Audiences? Testing for the Elusive Familiarity Backfire Effect,” Apr. 2020.
- [26] B. Nyhan and J. Reifler, “When Corrections Fail: The Persistence of Political Misperceptions,” *Political Behavior*, vol. 32, no. 2, pp. 303–330, Jun. 2010.
- [27] N. J. Conroy, V. L. Rubin, and Y. Chen, “Automatic deception detection: Methods for finding fake news,” *Proc. Assoc. Info. Sci. Tech.*, vol. 52, no. 1, pp. 1–4, Feb. 2015.

- [28] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, “Automatic Detection of Fake News,” Aug. 2017.
- [29] V. L. Rubin, Y. Chen, and N. J. Conroy, “Deception detection for news: Three types of fakes,” *Proc. Assoc. Info. Sci. Tech.*, vol. 52, no. 1, pp. 1–4, Feb. 2015.
- [30] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake News Detection on Social Media: A Data Mining Perspective,” Sep. 2017.
- [31] N. Ruchansky, S. Seo, and Y. Liu, “CSI: A Hybrid Deep Model for Fake News Detection,” in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, ser. CIKM ’17, New York, NY, USA: Association for Computing Machinery, Nov. 2017, pp. 797–806.
- [32] J. W. Kim, A. Guess, B. Nyhan, and J. Reifler, “The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity,” *Journal of Communication*, 2020.
- [33] Y. Hua, M. Naaman, and T. Ristenpart, “Characterizing Twitter Users Who Engage in Adversarial Interactions against Political Candidates,” en, in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu HI USA: ACM, Apr. 2020, pp. 1–13, ISBN: 978-1-4503-6708-0. DOI: [10.1145/3313831.3376548](https://doi.org/10.1145/3313831.3376548). URL: <https://dl.acm.org/doi/10.1145/3313831.3376548> (visited on 09/02/2021).
- [34] Y. Hua, T. Ristenpart, and M. Naaman, “Towards measuring adversarial twitter interactions against candidates in the us midterm elections,” in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, 2020, pp. 272–282.
- [35] C. A. Bail, L. P. Argyle, T. W. Brown, J. P. Bumpus, H. Chen, M. B. F. Hunzaker, J. Lee, M. Mann, F. Merhout, and A. Volfovsky, “Exposure to opposing views on social media can increase political polarization,” en, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 115, no. 37, pp. 9216–9221, Sep. 2018.
- [36] H. Allcott, L. Braghieri, S. Eichmeyer, and M. Gentzkow, “The welfare effects of social media,” *American Economic Review*, vol. 110, no. 3, pp. 629–76, 2020.
- [37] C. Sunstein and C. R. Sunstein, *# Republic*. Princeton university press, 2018.
- [38] E. Colleoni, A. Rozza, and A. Arvidsson, “Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data,” *Journal of communication*, vol. 64, no. 2, pp. 317–332, 2014.
- [39] P. Barberá, “Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data,” *Political analysis*, vol. 23, no. 1, pp. 76–91, 2015, Publisher: Cambridge University Press.
- [40] M. Mosleh, C. Martel, D. Eckles, and D. G. Rand, “Shared partisanship dramatically increases social tie formation in a Twitter field experiment,” *Proceedings of the National Academy of Sciences*, vol. 118, no. 7, 2021, Publisher: National Acad Sciences.
- [41] M. Gentzkow and J. M. Shapiro, “Ideological segregation online and offline,” *The Quarterly Journal of Economics*, vol. 126, no. 4, pp. 1799–1839, 2011.

- [42] A. M. Guess, “(almost) everything in moderation: New evidence on americans’ online media diets,” *American Journal of Political Science*, 2021.
- [43] 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.
- [44] 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.
- [45] 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*, pp. 1–17, 2021.
- [46] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer, “Fake news on Twitter during the 2016 U.S. presidential election,” *en, Science*, vol. 363, no. 6425, pp. 374–378, Jan. 2019.
- [47] 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, 2021.
- [48] X. Yu, M. Wojcieszak, and A. Casas, “Affective polarization on social media: In-party love among american politicians, greater engagement with out-party hate among ordinary users,” 2021.
- [49] N. Sirlin, Z. Epstein, A. A. Arechar, and D. G. Rand, “Digital literacy is associated with more discerning accuracy judgments but not sharing intentions,” *Harvard Kennedy School Misinformation Review*, 2021.
- [50] Z. Epstein, N. Sirlin, A. A. Arechar, G. Pennycook, and D. Rand, “Social media sharing reduces truth discernment,” 2021.
- [51] 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.
- [52] J. Allen, A. A. Arechar, G. Pennycook, and D. G. Rand, “Scaling up fact-checking using the wisdom of crowds,” *Preprint at <https://doi.org/10.31234/osf.io/9qdza>*, 2020.
- [53] P. Resnick, A. Alfayez, J. Im, and E. Gilbert, “Informed crowds can effectively identify misinformation,” *arXiv preprint arXiv:2108.07898*, 2021.
- [54] G. Pennycook and D. G. Rand, “Fighting misinformation on social media using crowd-sourced judgments of news source quality,” *Proceedings of the National Academy of Sciences*, vol. 116, no. 7, pp. 2521–2526, 2019.
- [55] N. Dias, G. Pennycook, and D. G. Rand, “Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media,” *Harvard Kennedy School Misinformation Review*, vol. 1, no. 1, 2020.

- [56] W. Godel, Z. Sanderson, K. Aslett, J. Nagler, R. Bonneau, N. Persily, and J. A. Tucker, “Moderating with the mob: Evaluating the efficacy of real-time crowdsourced fact-checking,” *Journal of Online Trust and Safety*, vol. 1, no. 1, 2021.
- [57] 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.
- [58] K. Coleman, *Introducing Birdwatch, a community-based approach to misinformation*, en_us, 2021. URL: https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.html (visited on 05/17/2021).
- [59] S. Iyengar, K. S. Hahn, J. A. Krosnick, and J. Walker, “Selective exposure to campaign communication: The role of anticipated agreement and issue public membership,” *The Journal of Politics*, vol. 70, no. 1, pp. 186–200, 2008.
- [60] S. Knobloch-Westerwick and J. Meng, “Looking the other way: Selective exposure to attitude-consistent and counterattitudinal political information,” *Communication Research*, vol. 36, no. 3, pp. 426–448, 2009.
- [61] N. J. Stroud, “Polarization and partisan selective exposure,” *Journal of communication*, vol. 60, no. 3, pp. 556–576, 2010.
- [62] A. Guess, B. Nyhan, B. Lyons, and J. Reifler, “Avoiding the echo chamber about echo chambers,” *Knight Foundation*, vol. 2, 2018.
- [63] S. Flaxman, S. Goel, and J. M. Rao, “Ideological segregation and the effects of social media on news consumption,” *Available at SSRN*, vol. 2363701, 2013.
- [64] M. Prior, “Media and political polarization,” *Annual Review of Political Science*, vol. 16, pp. 101–127, 2013.
- [65] 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.
- [66] S. Messing and S. J. Westwood, “Selective exposure in the age of social media: Endorsements trump partisan source affiliation when selecting news online,” *Communication research*, vol. 41, no. 8, pp. 1042–1063, 2014.
- [67] A. Boutet, H. Kim, and E. Yoneki, “What’s in twitter, i know what parties are popular and who you are supporting now!” *Social network analysis and mining*, vol. 3, no. 4, pp. 1379–1391, 2013.
- [68] M. D. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini, “Political polarization on twitter,” in *Fifth international AAAI conference on weblogs and social media*, 2011.
- [69] J. Shin and K. Thorson, “Partisan Selective Sharing: The Biased Diffusion of Fact-Checking Messages on Social Media: Sharing Fact-Checking Messages on Social Media,” en, *Journal of Communication*, vol. 67, no. 2, pp. 233–255, Apr. 2017, ISSN: 00219916. DOI: [10.1111/jcom.12284](https://doi.org/10.1111/jcom.12284). URL: <https://academic.oup.com/joc/article/67/2/233-255/4082394> (visited on 09/05/2021).

- [70] G. Pennycook, T. D. Cannon, and D. G. Rand, “Prior exposure increases perceived accuracy of fake news.,” *Journal of experimental psychology: general*, vol. 147, no. 12, p. 1865, 2018.
- [71] D. M. Kahan, “Ideology, motivated reasoning, and cognitive reflection: An experimental study,” *Judgment and Decision making*, vol. 8, pp. 407–24, 2012.
- [72] Z. Kunda, “The case for motivated reasoning.,” *Psychological bulletin*, vol. 108, no. 3, p. 480, 1990.
- [73] D. M. Kahan, “The politically motivated reasoning paradigm, part 1: What politically motivated reasoning is and how to measure it,” *Emerging trends in the social and behavioral sciences*, vol. 29, 2016.
- [74] C. S. Taber and M. Lodge, “Motivated skepticism in the evaluation of political beliefs,” *American journal of political science*, vol. 50, no. 3, pp. 755–769, 2006.
- [75] B. Nyhan and J. Reifler, “When corrections fail: The persistence of political misperceptions,” *Political Behavior*, vol. 32, no. 2, pp. 303–330, 2010.
- [76] T. Wood and E. Porter, “The elusive backfire effect: Mass attitudes’ steadfast factual adherence,” en, *Polit. Behav.*, vol. 41, no. 1, pp. 135–163, Mar. 2019, Publisher: Springer Science and Business Media LLC.
- [77] B. Swire-Thompson, J. DeGutis, and D. Lazer, “Searching for the Backfire Effect: Measurement and Design Considerations,” eng, *Journal of applied research in memory and cognition*, vol. 9, no. 3, pp. 286–299, Sep. 2020, Edition: 2020/09/02 Publisher: The Authors. Published by Elsevier Inc. on behalf of Society for Applied Research in Memory and Cognition., ISSN: 2211-3681. DOI: [10.1016/j.jarmac.2020.06.006](https://doi.org/10.1016/j.jarmac.2020.06.006). URL: <https://pubmed.ncbi.nlm.nih.gov/32905023>.
- [78] B. M. Tappin, G. Pennycook, and D. G. Rand, “Bayesian or biased? analytic thinking and political belief updating,” *Cognition*, vol. 204, p. 104375, 2020.
- [79] B. M. Tappin, G. Pennycook, and D. G. Rand, “Rethinking the link between cognitive sophistication and politically motivated reasoning.,” *Journal of Experimental Psychology: General*, 2020.
- [80] 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.
- [81] G. Pennycook and D. G. Rand, “The psychology of fake news,” *Trends in cognitive sciences*, 2021.
- [82] S. Mukerjee, K. Jaidka, and Y. Lelkes, “The political landscape of the us twitterverse,” 2020.
- [83] S. Wojcik and A. Hughes, “Sizing up twitter users,” *Pew Research Center*, vol. 24, 2019.
- [84] J. Becker, E. Porter, and D. Centola, “The wisdom of partisan crowds,” en, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 116, no. 22, pp. 10717–10722, May 2019, Publisher: National Acad Sciences.

- [85] 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.
- [86] F. Shi, M. Teplitskiy, E. Duede, and J. A. Evans, “The wisdom of polarized crowds,” *Nat Hum Behav*, vol. 3, no. 4, pp. 329–336, Apr. 2019, Publisher: nature.com.
- [87] J. Kim, B. Tabibian, A. Oh, B. Schölkopf, and M. Gomez-Rodriguez, “Leveraging the crowd to detect and reduce the spread of fake news and misinformation,” in *Proceedings of the eleventh ACM international conference on web search and data mining*, 2018, pp. 324–332.
- [88] S. Tschitschek, A. Singla, M. Gomez Rodriguez, A. Merchant, and A. Krause, “Fake news detection in social networks via crowd signals,” in *Companion Proceedings of the The Web Conference 2018*, 2018, pp. 517–524.
- [89] M. Siering, J. Muntermann, and B. Rajagopalan, “Explaining and predicting online review helpfulness: The role of content and reviewer-related signals,” *Decision Support Systems*, vol. 108, pp. 1–12, 2018.
- [90] G. O. Diaz and V. Ng, “Modeling and prediction of online product review helpfulness: A survey,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 698–708.
- [91] Z. Wang, S. Hale, D. I. Adelani, P. Grabowicz, T. Hartman, F. Flöck, and D. Jurgens, “Demographic inference and representative population estimates from multilingual social media data,” in *The World Wide Web Conference*, ACM, 2019, pp. 2056–2067.
- [92] R. Caruana and A. Niculescu-Mizil, “An empirical comparison of supervised learning algorithms,” in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 161–168.
- [93] M. Desai and M. A. Mehta, “Techniques for sentiment analysis of twitter data: A comprehensive survey,” in *2016 International Conference on Computing, Communication and Automation (ICCCA)*, 2016, pp. 149–154. DOI: [10.1109/CCAA.2016.7813707](https://doi.org/10.1109/CCAA.2016.7813707).
- [94] C. Silverman, “Here are 50 of the biggest fake news hits on facebook from 2016,” *Buzzfeed News*, pp. 1–12, 2016.
- [95] A. Eveleigh, C. Jennett, A. Blandford, P. Brohan, and A. L. Cox, “Designing for dabblers and deterring drop-outs in citizen science,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2014, pp. 2985–2994.
- [96] M. J. Raddick, G. Bracey, P. L. Gay, C. J. Lintott, P. Murray, K. Schawinski, A. S. Szalay, and J. Vandenberg, “Galaxy zoo: Exploring the motivations of citizen science volunteers,” *arXiv preprint arXiv:0909.2925*, 2009.
- [97] A. M. Land-Zandstra, J. L. Devilee, F. Snik, F. Buurmeijer, and J. M. van den Broek, “Citizen science on a smartphone: Participants’ motivations and learning,” *Public Understanding of Science*, vol. 25, no. 1, pp. 45–60, 2016.

- [98] L. Huddy, L. Mason, and L. Aarøe, “Expressive partisanship: Campaign involvement, political emotion, and partisan identity,” *American Political Science Review*, vol. 109, no. 1, pp. 1–17, 2015.
- [99] C. Avery, P. Resnick, and R. Zeckhauser, “The market for evaluations,” *American economic review*, vol. 89, no. 3, pp. 564–584, 1999.
- [100] J. A. Tucker, A. Guess, P. Barberá, C. Vaccari, A. Siegel, S. Sanovich, D. Stukal, and B. Nyhan, “Social media, political polarization, and political disinformation: A review of the scientific literature,” *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)*, 2018.
- [101] S. Iyengar, Y. Lelkes, M. Levendusky, N. Malhotra, and S. J. Westwood, “The origins and consequences of affective polarization in the united states,” *Annual Review of Political Science*, vol. 22, pp. 129–146, 2019.
- [102] 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,” en, in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama Japan: ACM, May 2021, pp. 1–13, ISBN: 978-1-4503-8096-6. DOI: [10.1145/3411764.3445642](https://doi.org/10.1145/3411764.3445642). URL: <https://dl.acm.org/doi/10.1145/3411764.3445642> (visited on 09/02/2021).
- [103] T. Yasseri and F. Menczer, “Can the wikipedia moderation model rescue the social marketplace of ideas?” *arXiv preprint arXiv:2104.13754*, 2021.
- [104] Birdwatch, *Today, we’re updating how notes are elevated in Birdwatch! This change will give more weight to contributors whose notes and ratings are consistently found helpful by others.* en, Tweet, Jun. 2021. URL: <https://twitter.com/birdwatch/status/1404519791394758657> (visited on 09/07/2021).
- [105] B. M. Tappin, G. Pennycook, and D. G. Rand, “Thinking clearly about causal inferences of politically motivated reasoning: Why paradigmatic study designs often undermine causal inference,” *Current Opinion in Behavioral Sciences*, vol. 34, pp. 81–87, 2020.
- [106] S. Atske, *Climate Change Remains Top Global Threat Across 19-Country Survey*, en-US, Aug. 2022. URL: <https://www.pewresearch.org/global/2022/08/31/climate-change-remains-top-global-threat-across-19-country-survey/> (visited on 05/13/2023).
- [107] D. M. J. Lazer, M. A. Baum, Y. Benkler, *et al.*, “The science of fake news,” en, *Science*, vol. 359, no. 6380, pp. 1094–1096, Mar. 2018.
- [108] C. Shao, G. L. Ciampaglia, A. Flammini, and F. Menczer, “Hoaxy: A platform for tracking online misinformation,” in *Proceedings of the 25th international conference companion on world wide web*, 2016, pp. 745–750.
- [109] C. Shao, P.-M. Hui, L. Wang, X. Jiang, A. Flammini, F. Menczer, and G. L. Ciampaglia, “Anatomy of an online misinformation network,” *Plos one*, vol. 13, no. 4, e0196087, 2018, Publisher: Public Library of Science San Francisco, CA USA.

- [110] S. Vosoughi, D. Roy, and S. Aral, “The spread of true and false news online,” en, *Science*, vol. 359, no. 6380, pp. 1146–1151, Mar. 2018.
- [111] S. Altay, R. K. Nielsen, and R. Fletcher, “Quantifying the “infodemic”: People turned to trustworthy news outlets during the 2020 coronavirus pandemic,” *Journal of Quantitative Description: Digital Media*, vol. 2, 2022.
- [112] G. Pennycook and D. G. Rand, “The psychology of fake news,” *Trends in cognitive sciences*, vol. 25, no. 5, pp. 388–402, 2021, Publisher: Elsevier.
- [113] S. Van der Linden, J. Roozenbeek, R. Maertens, M. Basol, O. Kácha, S. Rathje, and C. S. Traberg, “How can psychological science help counter the spread of fake news?” *The Spanish Journal of Psychology*, vol. 24, e25, 2021, Publisher: Cambridge University Press.
- [114] S. Van Der Linden, E. Maibach, J. Cook, A. Leiserowitz, and S. Lewandowsky, “Inoculating against misinformation,” *Science*, vol. 358, no. 6367, pp. 1141–1142, 2017, Publisher: American Association for the Advancement of Science.
- [115] S. Altay, M. Berriche, and A. Acerbi, “Misinformation on misinformation: Conceptual and methodological challenges,” *Social Media+ Society*, vol. 9, no. 1, 2023, Publisher: SAGE Publications Sage UK: London, England.
- [116] G. Murphy, C. de Saint Laurent, M. Reynolds, O. Aftab, K. Hegarty, Y. Sun, and C. M. Greene, “What do we study when we study misinformation? a scoping review of experimental research (2016-2022),” *Harvard Kennedy School Misinformation Review*, 2023.
- [117] S. Athey, K. Grabarz, M. Luca, and N. Wernerfelt, “Digital public health interventions at scale: The impact of social media advertising on beliefs and outcomes related to COVID vaccines,” *Proceedings of the National Academy of Sciences*, vol. 120, no. 5, e2208110120, 2023, Publisher: National Acad Sciences.
- [118] L. R. Baden, H. M. El Sahly, B. Essink, *et al.*, “Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine,” en, *N. Engl. J. Med.*, vol. 384, no. 5, pp. 403–416, Feb. 2021.
- [119] F. P. Polack, S. J. Thomas, N. Kitchin, *et al.*, “Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine,” en, *N. Engl. J. Med.*, vol. 383, no. 27, pp. 2603–2615, Dec. 2020.
- [120] J. Holder, *Tracking Coronavirus Vaccinations Around the World*, Mar. 2023. URL: <https://www.nytimes.com/interactive/2021/world/covid-vaccinations-tracker.html> (visited on 05/10/2023).
- [121] N. Bose and E. Culliford, *Biden says Facebook, others ‘killing people’ by carrying COVID misinformation* | Reuters, Jul. 2021. URL: <https://www.reuters.com/business/healthcare-pharmaceuticals/white-house-says-facebooks-steps-stop-vaccine-misinformation-are-inadequate-2021-07-16/> (visited on 05/01/2023).

- [122] J. Aw, J. J. B. Seng, S. S. Y. Seah, and L. L. Low, “COVID-19 Vaccine Hesitancy—A Scoping Review of Literature in High-Income Countries,” en, *Vaccines*, vol. 9, no. 8, p. 900, Aug. 2021, Number: 8 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2076-393X. DOI: [10.3390/vaccines9080900](https://doi.org/10.3390/vaccines9080900). URL: <https://www.mdpi.com/2076-393X/9/8/900> (visited on 05/10/2023).
- [123] A. Bridgman, E. Merkley, P. J. Loewen, T. Owen, D. Ruths, L. Teichmann, and O. Zhilin, “The causes and consequences of COVID-19 misperceptions: Understanding the role of news and social media,” en-US, *Harvard Kennedy School Misinformation Review*, vol. 1, no. 3, Jun. 2020. DOI: [10.37016/mr-2020-028](https://doi.org/10.37016/mr-2020-028). URL: <https://misinformreview.hks.harvard.edu/article/the-causes-and-consequences-of-covid-19-misperceptions-understanding-the-role-of-news-and-social-media/> (visited on 05/10/2023).
- [124] N. Puri, E. A. Coomes, H. Haghbayan, and K. Gunaratne, “Social media and vaccine hesitancy: New updates for the era of COVID-19 and globalized infectious diseases,” *Human Vaccines & Immunotherapeutics*, vol. 16, no. 11, pp. 2586–2593, Nov. 2020, Publisher: Taylor & Francis _eprint: <https://doi.org/10.1080/21645515.2020.1780846>, ISSN: 2164-5515. DOI: [10.1080/21645515.2020.1780846](https://doi.org/10.1080/21645515.2020.1780846). URL: <https://doi.org/10.1080/21645515.2020.1780846> (visited on 05/10/2023).
- [125] S. van der Linden, “Misinformation: Susceptibility, spread, and interventions to immunize the public,” en, *Nature Medicine*, vol. 28, no. 3, pp. 460–467, Mar. 2022, Number: 3 Publisher: Nature Publishing Group, ISSN: 1546-170X. DOI: [10.1038/s41591-022-01713-6](https://doi.org/10.1038/s41591-022-01713-6). URL: <https://www.nature.com/articles/s41591-022-01713-6> (visited on 05/11/2023).
- [126] S. van der Linden, *We need a gold standard for randomised control trials studying misinformation and vaccine hesitancy on social media*, 2023.
- [127] C. de Saint Laurent, G. Murphy, K. Hegarty, and C. M. Greene, “Measuring the effects of misinformation exposure and beliefs on behavioural intentions: A COVID-19 vaccination study,” *Cognitive Research: Principles and Implications*, vol. 7, no. 1, p. 87, 2022, Publisher: Springer.
- [128] 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,” en, *Nat Hum Behav*, Feb. 2021.
- [129] A. M. Guess, B. Nyhan, Z. O’Keeffe, and J. Reifler, “The sources and correlates of exposure to vaccine-related (mis) information online,” *Vaccine*, vol. 38, no. 49, pp. 7799–7805, 2020.
- [130] M. R. DeVerna, F. Pierri, B. T. Truong, J. Bollenbacher, D. Axelrod, N. Loynes, C. Torres-Lugo, K.-C. Yang, F. Menczer, and J. Bryden, “CoVaxxy: A collection of English-language Twitter posts about COVID-19 vaccines,” in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 15, 2021, pp. 992–999.
- [131] T. Hossain, R. L. Logan IV, A. Ugarte, Y. Matsubara, S. Young, and S. Singh, “COVIDLies: Detecting COVID-19 misinformation on social media,” in *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, 2020.

- [132] E. Porter, T. J. Wood, and Y. Velez, “Correcting COVID-19 vaccine misinformation in Ten countries,” 2022, Publisher: OSF Preprints.
- [133] H. H. Clark, “The language-as-fixed-effect fallacy: A critique of language statistics in psychological research,” en, *Journal of Verbal Learning and Verbal Behavior*, vol. 12, no. 4, pp. 335–359, Aug. 1973, ISSN: 00225371. DOI: [10.1016/S0022-5371\(73\)80014-3](https://doi.org/10.1016/S0022-5371(73)80014-3). URL: <https://linkinghub.elsevier.com/retrieve/pii/S0022537173800143> (visited on 01/16/2024).
- [134] T. Yarkoni, “The generalizability crisis,” *Behavioral and Brain Sciences*, vol. 45, e1, 2022, Publisher: Cambridge University Press.
- [135] P. Patwa, S. Sharma, S. Pykl, V. Guptha, G. Kumari, M. S. Akhtar, A. Ekbal, A. Das, and T. Chakraborty, “Fighting an infodemic: Covid-19 fake news dataset,” in *Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1*, Springer, 2021, pp. 21–29.
- [136] M. Parks, “Few facts, millions of clicks: Fearmongering vaccine stories go viral online,” *NPR. March*, vol. 25, 2021.
- [137] I. J. Borges do Nascimento, A. B. Pizarro, J. M. Almeida, N. Azzopardi-Muscat, M. A. Gonçalves, M. Björklund, and D. Novillo-Ortiz, “Infodemics and health misinformation: A systematic review of reviews,” *Bulletin of the World Health Organization*, vol. 100, no. 9, pp. 544–561, Sep. 2022, ISSN: 0042-9686. DOI: [10.2471/BLT.21.287654](https://doi.org/10.2471/BLT.21.287654). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9421549/> (visited on 05/12/2023).
- [138] E. Chen, J. Jiang, H.-C. H. Chang, G. Muric, and E. Ferrara, “Charting the Information and Misinformation Landscape to Characterize Misinfodemics on Social Media: COVID-19 Infodemiology Study at a Planetary Scale,” EN, *JMIR Infodemiology*, vol. 2, no. 1, e32378, Feb. 2022, Company: JMIR Infodemiology Distributor: JMIR Infodemiology Institution: JMIR Infodemiology Label: JMIR Infodemiology Publisher: JMIR Publications Inc., Toronto, Canada. DOI: [10.2196/32378](https://doi.org/10.2196/32378). URL: <https://infodemiology.jmir.org/2022/1/e32378> (visited on 05/12/2023).
- [139] M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoli, A. L. Schmidt, P. Zola, F. Zollo, and A. Scala, “The COVID-19 social media infodemic,” *Scientific reports*, vol. 10, no. 1, pp. 1–10, 2020, Publisher: Springer.
- [140] F. Pierri, M. R. DeVerna, K.-C. Yang, D. Axelrod, J. Bryden, and F. Menczer, “One Year of COVID-19 Vaccine Misinformation on Twitter: Longitudinal Study,” EN, *Journal of Medical Internet Research*, vol. 25, no. 1, e42227, Feb. 2023, Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada. DOI: [10.2196/42227](https://doi.org/10.2196/42227). URL: <https://www.jmir.org/2023/1/e42227> (visited on 05/08/2023).
- [141] S. Messing, C. DeGregorio, B. Hillenbrand, G. King, S. Mahanti, C. Nayak, N. Persily, State, Bogdan, and A. Wilkins, *Facebook Privacy-Protected Full URLs Data Set*. 2020.

- [142] K. L. Milkman, L. Gandhi, M. S. Patel, H. N. Graci, D. M. Gromet, H. Ho, J. S. Kay, T. W. Lee, J. Rothschild, and J. E. Bogard, “A 680,000-person megastudy of nudges to encourage vaccination in pharmacies,” *Proceedings of the National Academy of Sciences*, vol. 119, no. 6, e2115126119, 2022, Publisher: National Acad Sciences.
- [143] M. Srivastava, T. Hashimoto, and P. Liang, “Robustness to Spurious Correlations via Human Annotations,” en, in *Proceedings of the 37th International Conference on Machine Learning*, ISSN: 2640-3498, PMLR, Nov. 2020, pp. 9109–9119. URL: <https://proceedings.mlr.press/v119/srivastava20a.html> (visited on 05/29/2023).
- [144] C. Dwork, “Differential privacy,” in *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33*, Springer, 2006, pp. 1–12.
- [145] A. Guess, K. Aslett, J. Tucker, R. Bonneau, and J. Nagler, “Cracking open the news feed: Exploring what us Facebook users see and share with large-scale platform data,” *Journal of Quantitative Description: Digital Media*, vol. 1, 2021.
- [146] J. Allen, M. Mobius, D. M. Rothschild, and D. J. Watts, “Research note: Examining potential bias in large-scale censored data,” *Harvard Kennedy School Misinformation Review*, 2021.
- [147] J. Lasser, S. T. Aroyehun, Almog Simchon, F. Carrella, D. Garcia, and S. Lewandowsky, “Social media sharing of low-quality news sources by political elites,” *PNAS nexus*, vol. 1, no. 4, pgac186, 2022, Publisher: Oxford University Press.
- [148] L. Hewitt, D. Broockman, A. Coppock, B. M. Tappin, J. Slezak, N. Lubin, and M. Hamidian, “How experiments help campaigns persuade voters: Evidence from a large archive of campaigns’ own experiments,” *American Journal of Political Science*, vol. Forthcoming, 2023.
- [149] M. Borenstein, J. P. T. Higgins, L. V. Hedges, and H. R. Rothstein, “Basics of meta-analysis: I2 is not an absolute measure of heterogeneity,” en, *Research Synthesis Methods*, vol. 8, no. 1, pp. 5–18, 2017, ISSN: 1759-2887. DOI: [10.1002/jrsm.1230](https://doi.org/10.1002/jrsm.1230). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/jrsm.1230> (visited on 07/24/2023).
- [150] O. Sagi and L. Rokach, “Ensemble learning: A survey,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, e1249, 2018, Publisher: Wiley Online Library.
- [151] M. Müller, M. Salathé, and P. E. Kummervold, “Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter,” *arXiv preprint*, 2020.
- [152] *U.S.: Facebook users 2018-2027*, en. URL: <https://www.statista.com/statistics/408971/number-of-us-facebook-users/> (visited on 05/13/2023).
- [153] A. Coppock, “Persuasion in parallel,” in *Persuasion in Parallel*, University of Chicago Press, 2022.
- [154] J. B. Ruiz and R. A. Bell, “Predictors of intention to vaccinate against COVID-19: Results of a nationwide survey,” en, *Vaccine*, vol. 39, no. 7, pp. 1080–1086, Feb. 2021, ISSN: 0264-410X. DOI: [10.1016/j.vaccine.2021.01.010](https://doi.org/10.1016/j.vaccine.2021.01.010). URL: <https://www.sciencedirect.com/science/article/pii/S0264410X21000141> (visited on 05/01/2023).

- [155] A. M. Guess, B. Nyhan, Z. O’Keeffe, and J. Reifler, “The sources and correlates of exposure to vaccine-related (mis)information online,” en, *Vaccine*, vol. 38, no. 49, pp. 7799–7805, Nov. 2020, ISSN: 0264-410X. DOI: [10.1016/j.vaccine.2020.10.018](https://doi.org/10.1016/j.vaccine.2020.10.018). URL: <https://www.sciencedirect.com/science/article/pii/S0264410X20313116> (visited on 05/10/2023).
- [156] S. Loomba, R. Maertens, J. Roozenbeek, F. M. Götz, S. van der Linden, and A. De Figueiredo, “Ability to detect fake news predicts sub-national variation in COVID-19 vaccine uptake across the UK,” *medRxiv*, pp. 2023–05, 2023, Publisher: Cold Spring Harbor Laboratory Press. URL: <https://www.medrxiv.org/content/10.1101/2023.05.10.23289764.abstract> (visited on 01/16/2024).
- [157] K. Aslett, A. M. Guess, R. Bonneau, J. Nagler, and J. A. Tucker, “News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions,” *Science advances*, vol. 8, no. 18, eabl3844, 2022, Publisher: American Association for the Advancement of Science.
- [158] K. Clayton, S. Blair, J. A. Busam, S. Forstner, J. Glance, G. Green, A. Kawata, A. Kovvuri, J. Martin, and E. Morgan, “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, Publisher: Springer.
- [159] A. M. Guess, M. Lerner, B. Lyons, J. M. Montgomery, B. Nyhan, J. Reifler, and N. Sircar, “A digital media literacy intervention increases discernment between mainstream and false news in the United States and India,” EN, *Proceedings of the National Academy of Sciences*, vol. 117, no. 27, pp. 15 536–15 545, Jul. 2020, Company: National Academy of Sciences Distributor: National Academy of Sciences ISBN: 9781920498115 Institution: National Academy of Sciences Label: National Academy of Sciences Publisher: Proceedings of the National Academy of Sciences. DOI: [10.1073/pnas.1920498117](https://doi.org/10.1073/pnas.1920498117). URL: <https://www.pnas.org/doi/abs/10.1073/pnas.1920498117> (visited on 12/16/2022).
- [160] K. Hartwig, F. Doell, and C. Reuter, *The Landscape of User-centered Misinformation Interventions – A Systematic Literature Review*, arXiv:2301.06517 [cs], Jan. 2023. DOI: [10.48550/arXiv.2301.06517](https://doi.org/10.48550/arXiv.2301.06517). URL: <http://arxiv.org/abs/2301.06517> (visited on 05/01/2023).
- [161] *NewsGuard - Combating Misinformation with Trust Ratings for News*, en-US. URL: <https://www.newsguardtech.com/> (visited on 05/01/2023).
- [162] A. Oeldorf-Hirsch, M. Schmierbach, A. Appelman, and M. P. Boyle, “The Ineffectiveness of Fact-Checking Labels on News Memes and Articles,” *Mass Communication and Society*, vol. 23, no. 5, pp. 682–704, Sep. 2020, Publisher: Routledge _eprint: <https://doi.org/10.1080/15205436.2020.1733613>, ISSN: 1520-5436. DOI: [10.1080/15205436.2020.1733613](https://doi.org/10.1080/15205436.2020.1733613). URL: <https://doi.org/10.1080/15205436.2020.1733613> (visited on 05/01/2023).
- [163] S. Shabani and M. Sokhn, “Hybrid Machine-Crowd Approach for Fake News Detection,” in *2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC)*, Oct. 2018, pp. 299–306.

- [164] C. Fong and J. Grimmer, “Causal inference with latent treatments,” *American Journal of Political Science*, 2021, Publisher: Wiley Online Library.
- [165] M. H. Graham and A. Coppock, “Asking About Attitude Change,” *Public Opinion Quarterly*, vol. 85, no. 1, pp. 28–53, Jul. 2021, ISSN: 0033-362X. DOI: [10.1093/poq/nfab009](https://doi.org/10.1093/poq/nfab009). URL: <https://doi.org/10.1093/poq/nfab009> (visited on 05/29/2023).
- [166] J. Cheng and M. S. Bernstein, “Flock: Hybrid Crowd-Machine Learning Classifiers,” in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, ser. CSCW ’15, New York, NY, USA: Association for Computing Machinery, Feb. 2015, pp. 600–611, ISBN: 978-1-4503-2922-4. DOI: [10.1145/2675133.2675214](https://doi.org/10.1145/2675133.2675214). URL: <https://dl.acm.org/doi/10.1145/2675133.2675214> (visited on 05/30/2023).
- [167] M. Groh, Z. Epstein, C. Firestone, and R. Picard, “Deepfake detection by human crowds, machines, and machine-informed crowds,” *Proceedings of the National Academy of Sciences*, vol. 119, no. 1, e2110013119, 2022, Publisher: National Acad Sciences.
- [168] R. J. Barro, “Vaccination rates and COVID outcomes across US states,” *Economics & Human Biology*, vol. 47, p. 101 201, 2022, Publisher: Elsevier.
- [169] M. Hlavac, “Stargazer: Well-formatted regression and summary statistics tables,” 2018, R package version 5.2.2. URL: <https://CRAN.R-project.org/package=stargazer>.
- [170] A. Gerber and D. Green, *Field Experiments: Design, Analysis, and Interpretation*. New York: W. W, 2012.
- [171] S. Wager and S. Athey, “Estimation and Inference of Heterogeneous Treatment Effects using Random Forests,” en, *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1228–1242, Jul. 2018, ISSN: 0162-1459, 1537-274X. DOI: [10.1080/01621459.2017.1319839](https://doi.org/10.1080/01621459.2017.1319839). URL: <https://www.tandfonline.com/doi/full/10.1080/01621459.2017.1319839> (visited on 01/16/2024).
- [172] S. Athey and S. Wager, “Estimating treatment effects with causal forests: An application,” *Observational studies*, vol. 5, no. 2, pp. 37–51, 2019, Publisher: University of Pennsylvania Press. URL: <https://muse.jhu.edu/pub/56/article/793356/summary> (visited on 01/16/2024).
- [173] J. G. Bullock, “Partisan bias and the Bayesian ideal in the study of public opinion,” *The Journal of Politics*, vol. 71, no. 3, pp. 1109–1124, 2009, Publisher: Cambridge University Press New York, USA.
- [174] E. Kamenica and M. Gentzkow, “Bayesian persuasion,” *American Economic Review*, vol. 101, no. 6, pp. 2590–2615, 2011, Publisher: American Economic Association.
- [175] D. Broockman and J. Kalla, “When and Why Are Campaigns’ Persuasive Effects Small? Evidence from the 2020 US Presidential Election,” 2020.
- [176] M. Harrer, P. Cuijpers, T. A. Furukawa, and D. D. Ebert, *Doing meta-analysis with R: A hands-on guide*. CRC press, 2021.
- [177] K. P. Burnham and D. R. Anderson, “Multimodel inference: Understanding AIC and BIC in model selection,” *Sociological methods & research*, vol. 33, no. 2, pp. 261–304, 2004, Publisher: Sage Publications Sage CA: Thousand Oaks, CA.