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

Jennifer Allen[1], Duncan J. Watts[2,3,4], David G. Rand[1,5,6]

[1]Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA;
[2]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA;
[3]Annenberg School for Communication, University of Pennsylvania, Philadelphia, PA;
[4]Operations, Information, and Decisions Department, University of Pennsylvania, Philadelphia, PA;
[5]Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA;
[6]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge MA

Researchers and public health officials have attributed the low uptake of the COVID-19 vaccine in the US to misinformation spreading on social media. To evaluate this claim, we introduce a novel generalization framework, combining lab experiments, crowdsourcing, and machine learning, to estimate the causal effects on vaccination intentions of 13,206 vaccine-related URLs shared on Facebook during the initial vaccine rollout. Our model predicts that content that expressed skepticism about vaccine safety lowered vaccination intentions by approximately -2.33pp (95% QI: -2.28, -.328) per US Facebook user. Strikingly, content that was not flagged by fact-checkers but was vaccine-skeptical–mostly mainstream media articles that selectively reported on deaths following vaccination—was 50X more impactful in aggregate than outright misinformation. Although both our lab experiments and model show that when viewed, misinformation was significantly more harmful than factually accurate content, vaccine-skeptical mainstream media content accounted for vastly more views than misinformation on Facebook. While our work suggests that limiting the spread of misinformation has important public health benefits, it also highlights the potentially harmful effects of factually accurate but nonetheless misleading content published by mainstream sources.

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.[1] This attention is largely motivated by the assumption that misinformation causes substantial real-world harm–an assumption that is often justified by associations between misinformation on social media and events such as the unexpected election of Donald Trump in 2016, 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,[2–6] the prevalence of fake news,[7–10] and the cognitive psychology driving sharing of and belief in falsehoods,[1,11–14] consideration of the real-world *causal* impact of misinformation has been largely relegated to assertions in introductory paragraphs and discussion sections.[15]

This gap is particularly relevant in the context of COVID-19 vaccine misinformation: although the "infodemic" of viral falsehoods is frequently cited as an obstacle to the adoption of public health measures – e.g. President Joe Biden claimed Facebook was "killing people" by letting anti-vaccine misinformation spread on the platform, and numerous studies have examined the correlational relationship between social media use, endorsement of vaccine conspiracies, and vaccine hesitancy,[16–18] little work has been done to show a causal connection.[19] The few lab studies testing for a causal relationship between vaccine misinformation and behavioral intentions have shown conflicting

evidence,[20,21] and some research has suggested that pre-existing vaccine-skepticism inspires misinformation consumption, rather than vice-versa.[22] Thus, whether and to what extent misinformation has actually had an important impact on society remains an open question. Moreover, it is also possible 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.[22]

Here, we address this critical - but neglected - issue by introducing a framework for estimating causal impact at scale, and applying this approach to quantify the harm caused by COVID-19 vaccine misinformation on Facebook. We begin by asking what would be necessary for online misinformation to have the sweeping societal impact so broadly ascribed to it. We posit that for any type of information to have widespread impact on people's behavior, it must meet two criteria. First, it must influence behavior, conditional on being seen. Second, a large number of people must see it. Thus, we define impact as the product of exposure and persuasive influence: harmful misinformation which no one sees is not impactful, nor is misinformation that is widely seen but irrelevant to people's decision-making.

Our approach combines (i) results from experiments measuring the effect of different vaccine-related headlines on vaccination intentions with (ii) data about the exposure to vaccine-related URLs on Facebook. Generalizing from these results using a combination of crowdsourcing and machine learning, we then estimate the overall impact of vaccine-related content shared on Facebook on vaccination rates in the US. Critically, we model the impact of *all* vaccine-related content on Facebook, not only content that is demonstrably false. The reason is that, in contrast with prior work which treats inaccuracy as a proxy for harm, we directly measure harm as decreased willingness to take a COVID vaccine. 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.

Harm when exposed to COVID-19 vaccine (mis)information

First, we consider which types of vaccine content changed willingness to take a COVID-19 vaccine, conditional on exposure, using two large-scale 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-fallacy."[23,24] 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.[25] Furthermore, factually-accurate content might also increase vaccine hesitancy, e.g. news of the government pausing rollout of the Johnson & Johnson COVID-19 vaccine following concerns about blood clot risks.[26]

In Study 1, N=8603 American participants from Lucid were shown a neutral control headline or a single piece of vaccine misinformation from a set of 40 articles, videos, and posts previously debunked by fact-checkers. To assess impact, participants answered a set of questions regarding their willingness to take a COVID-19 vaccine (which we combine into a COVID-19 vaccination index) before and after exposure. Consistent with popular narratives about misinformation, we find that exposure to a single

piece of vaccine misinformation decreased vaccination intentions by 1.5 percentage points on average (p=.00004). This effect did not vary significantly based on participants' pre-treatment vaccination intentions, gender, age, political party, or vaccine status (p>.2 for all after Benjamini-Hochberg correction; see S1.5.4 for details).

We did, however, find substantially variation in effect size across different pieces of misinformation. A multi-level model finds that the standard deviation in effect size across content is 0.89pp, 60% the size of the overall treatment effect. This estimate suggests that the worst 10% of misinformation items had double the average effect – lowering vaccination intentions by 3% – while the bottom 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 Study 2, we further investigate this content-level heterogeneity. Rather than focusing only on debunked claims, we collected a representative set of 90 highly-engaging, vaccine-related articles sampled from Facebook, balanced across topic and domain quality. We then recruited N=10,122 American participants from Lucid and measured the causal effect of each piece of content on vaccination intentions using the same procedure as Study 1. (We sampled a large and varied set of content to precisely estimate which *dimensions* of content increase or decrease willingness to get a vaccine on average - as opposed to seeking to precisely estimate treatment effects for a small number of headlines).

To quantify relevant content dimensions, we presented a new set of raters with headlines from the 130 pieces of content collected in Studies 1 and 2, and had them label the headlines on whether they were 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 predicting the treatment effect of each headline on the vaccination intentions index using these five headline-level features as independent variables (here we present results pooling across Studies 1 and 2 for maximum power; see S1.5.3 for disaggregated analyses).

We find that the only content dimension that consistently predicts a headline's effect on vaccination intentions is the extent to which the headline suggests that the vaccine is harmful to a person's health (Figure 1): A 1 scale point increase in the headline claiming the vaccine is harmful to health was associated with an effect on vaccination intentions of -0.69pp (SE:0.19, p = .0003) for a model with just harmful-to-health as a predictor, and -0.49pp (SE: 0.23, p = .036) for a model including other potential moderating variables. (See S1.5.3 for associations with other content dimensions.)

Interestingly, we find no significant effect on vaccination intentions of whether the headline came from a low-quality vs. mainstream domain (β = -0.26, SE: .23, p = .25). Conversely, veracity (as judged by professional fact-checkers) *is* associated with a more negative effect on vaccination intentions (β = -0.85, SE: .27, p = .002). and false claims are more likely to suggest that the vaccine is harmful to health (β = .76, SE: .08, p<.00001, as can be seen in Figure 1). Critically, however, when predicting treatment effect size using both veracity and the extent to which the headline suggested the vaccine was harmful to health, as well as their interaction (with variables z-scored for interpretability), harmful to health remained significant (β=-.38, p=.005), while veracity did not (β=-.21, p=.17) (there was also no significant interaction, β=-.19., p = .16). These results indicate that suggesting the vaccine was harmful to health reduced vaccination intentions regardless of whether or not the headline was factually inaccurate. Conversely, we do not find strong evidence that falsehoods that do not also suggest harm are themselves harmful.

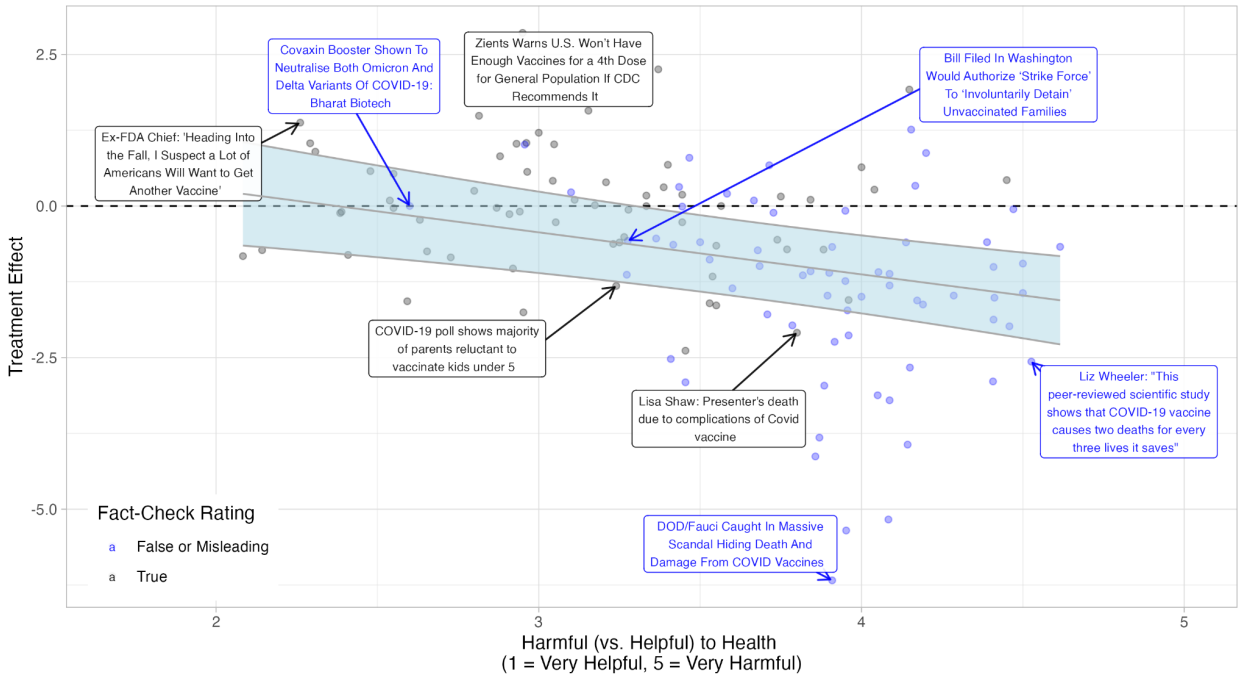

Figure 1: Effect on vaccine intent as a function of perceived harm for false/misleading (indicated in blue) and factually accurate articles (indicated in black). Overlaid in gray is the best-fit line and 95% confidence interval from a random-effects meta-regression with treatment effect as the outcome variable, the extent to which the article implied that the vaccine was harmful to a person's health as a moderator, and random effects for article and experiment.

Exposure to COVID-19 vaccine (mis)information on Facebook

We now examine levels of exposure to vaccine-related content on Facebook. Some prior 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.[27–30] However, other work has shown that fake news sharing and consumption is comparatively infrequent and highly concentrated among the heaviest news consumers,[7,9,10] even in the context of COVID-19.[8] 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 the large-scale Social Science One dataset released by Facebook to measure the actual views received by individual URLs on Facebook.[31] Specifically, we identify 13,206 URLs about the COVID-19 vaccine shared publicly more than 100 times on Facebook and published during the first 3 months of 2021 (the initial rollout period for the vaccine in the US).

We find that URLs flagged by professional fact-checkers as false, out-of-context, or mixture – which we will refer to as "misinformation" in our subsequent analyses – received 8.7 million views, accounting for only 0.3% of the 2.7 billion URL views during this time period (Figure 2A). Similarly, content from domains rated as low-credibility (based on Lasser et al, 2022) received only 5.1% of views.

Thus, exposure to outright misinformation about vaccines on Facebook was relatively infrequent, due to some combination of low baseline user viewership and explicit downranking by Facebook.).[10,32–35]

As shown in our survey experiments, however, even factually accurate content may still have had negative effects on vaccination intentions. For example, examining the top 10 most-viewed vaccine-related story clusters in the dataset reveals that several factually accurate articles published by mainstream news organizations cast doubt on the safety and efficacy of the vaccine (we will refer to content that is not misinformation but still casts doubt on the vaccine as "vaccine-skeptical"[22]). 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." URLs related to this story were viewed over 65 million times on Facebook – more than 20% of Facebook's US user base, and more than 7 times the number of views on all misinformation combined. We emphasize that this story, and others like it, were factually accurate and in many cases indicated the uncertainty surrounding the true cause of death. Nonetheless, the story's clear implication (especially from the "click-bait" style headline) was that the vaccine may be harmful to health.

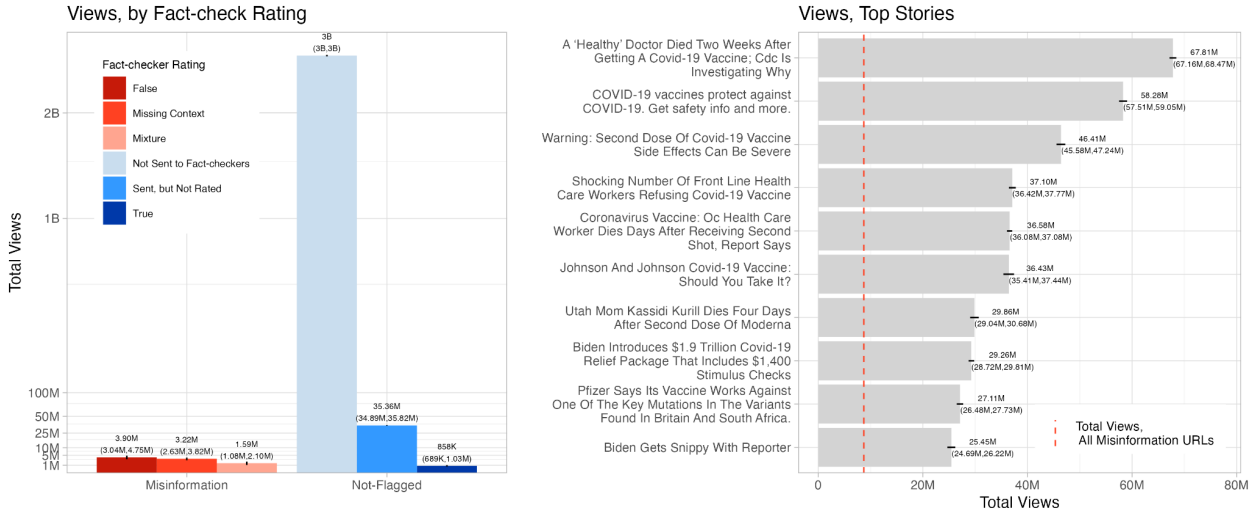

Figure 2: Exposure to vaccine related content on Facebook shared greater than 100 times on Facebook during the first 3 months of 2021. Panel A shows the total views for misinformation vs. non-misinformation content, broken down by fact-checker rating. The y-axis is square-root scaled for better visualization 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 story clusters across all content, where story clusters are composed of similar URLs that have been clustered together based on their headlines and descriptions (see Methods for more details).The aggregate number of views on all misinformation URLs is indicated by the red dashed line.

Scaling up estimates of harm conditional on exposure

To estimate the impact of misinformation and vaccine-skeptical content on Facebook, we need to combine the exposure data in the previous section with estimates of the causal effect of exposure for each headline. To generate such estimates, we use a combination of crowdsourcing and machine learning to generalize

the results from our survey experiments to the full Facebook URL dataset.

First, we recruited crowd raters from Cloud Research's Amazon Mechanical Turk panel and had them predict whether the 130 headlines from Studies 1 and 2 would cause people to be more or less likely to take a COVID-19 vaccine. We then create a "Crowdsourced Aggregate Score" by averaging this measure with the previously discussed crowdsourced ratings of (i) whether the headline suggests vaccines are harmful or helpful to health and (ii) headline accuracy (which, consistent with prior work,[36–38] agreed highly with expert judgments, r = .72, p <.000001).

The results are striking. Using random effects meta-regression to predict each of the 130 headlines' observed causal effect on vaccination intentions in Studies 1 and 2, we find that this crowdsourced aggregate score is highly predictive of the actual treatment effects (B=.84, p=.0001, pseudo-R^2=.74; Figure 3). Furthermore, the I^2, the percent of unexplained variance not attributable to sampling variation, of this model is 13.4, suggesting that there is little additional heterogeneity in treatment effects to be explained. 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 in the measurement of the treatment effects in Studies 1 and 2), it can successfully predict the expected *average* treatment effects across the range of crowdsourced predictions with high accuracy. 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 demonstrate the power of crowdsourcing for estimating causal effects.

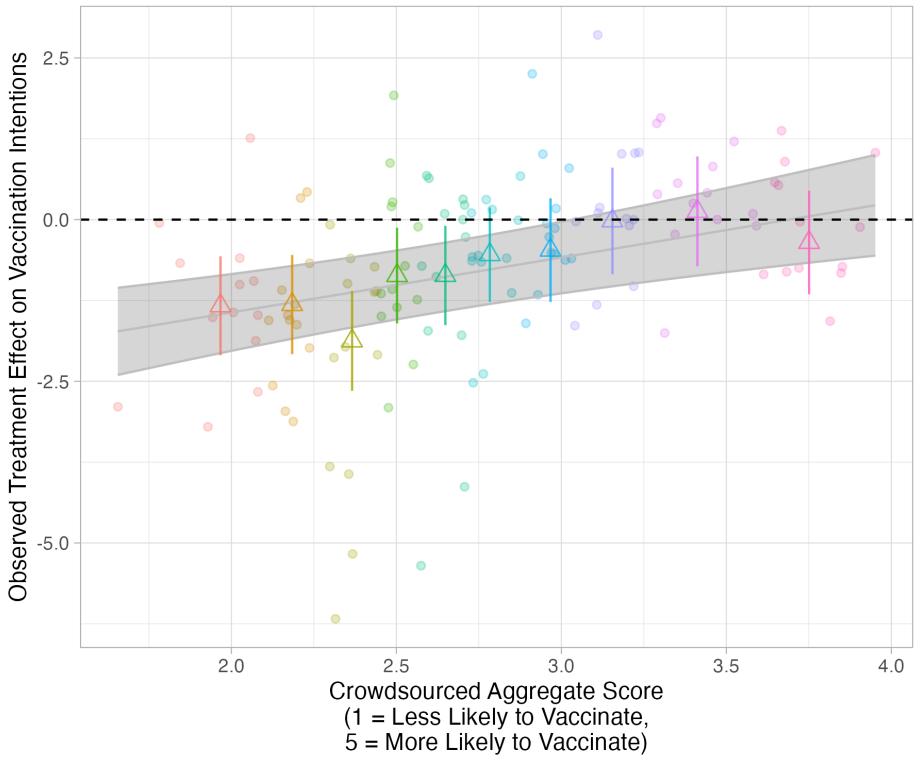

Figure 3: Treatment effect on vaccination intentions as a function of the Crowdsource Aggregate Score. Each point corresponds to one of the 130 items in Studies 1 and 2 and are colored by decile. The overlaid gray line is the best-fit line and 95% confidence-interval from a random effects meta-regression 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.

We then recruited additional raters and had them rate a random subset of 1200 of the 13,206 URLs from Facebook (upweighting URLs that were highly viewed, covering diverse events, and flagged by fact-checkers). We randomly split our labeled data into a 85/15 train-test split on our labeled data, stratified on whether or not the crowd thought the URL would increase or decrease vaccination intentions. We then trained a machine learning model using a Covid-Twitter-BERT architecture[39] to predict the crowdsourced scores for the full set of 13,206 URLs. We find that our model is highly accurate at predicting the crowdsourced aggregate score in our holdout test set; 86% of predicted aggregate scores were within half of a scale-point of the true aggregate score, and 99% were within 1 scale-point. On a binary classification task predicting whether the URL was below the scale midpoint, 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 URLs the crowd thought would increase support for vaccines hesitancy were predicted by our model to decrease support). (See S3.1 for details and robustness checks). To predict the causal effect (conditional on exposure) of each URL, our ultimate goal, we pass these predicted Crowdsourced Aggregate Scores into our meta-regression model to generate estimated treatment effects (i.e. effect of exposure) for the full set of URLs.

Consistent with the experimental results of Studies 1 and 2, we find that the estimated effect on vaccination intentions of the average misinformation URL is substantially more negative than the average URL not flagged by fact-checkers ($t(197)=-40.1$, $p<.00001$); Figure 4A). For example, the median misinformation URL has a estimated treatment effect of -1.34pp (95% CI: [-1.91,-.73]), more than four times larger in magnitude than the estimated treatment effect of the median non-flagged URL (-0.3pp, 95% CI: [-.92,.31]).

Quantifying harm caused by vaccine-skeptical (mis)information on Facebook

Does this observation imply that misinformation played an outsized role in causing harm on Facebook during the vaccine rollout? To answer this question, we must combine the estimated treatment effects shown in Figure 6a with the exposure data from the Social Science One dataset. As described above, misinformation URLs received only a small percentage of the total views - 98% of the over 500 million views of content estimated to reduce vaccinations was merely vaccine-skeptical rather than misinformation.

Thus, when we weight the estimated treatment effect for each URL by the number of views that URL received (Figure 4B), we see a very different picture: The impact of the misinformation posts is dwarfed by the impact of posts that were not flagged by fact-checkers but were vaccine-skeptical - posts which had somewhat less impact conditional on exposure than misinformation posts, but which were viewed by vastly more people. Thus, we take the product of exposure and estimated impact for each URL (and normalize by the total number of US Facebook users for interpretability). Doing so, we find that content that was not flagged by fact-checkers but was predicted to be harmful (i.e. below the midpoint on our aggregate crowd score) lowered vaccination rates by an estimated -2.28pp (CI: -3.4, -.99) per US Facebook user, compared to an effect of only -0.05pp (CI: -.07, -.02) for misinformation that is predicted to be harmful (see Figure 4C).

To shed more light on the content that was most harmful, we examine the stories predicted to have the largest effect on vaccination intentions - none of which were flagged by fact-checkers (see SI Figure S12). We find that coverage of young, healthy people's deaths following the vaccine - with headlines that did not contextualize how exceedingly rare such deaths were - achieved disproportionate reach, and therefore had disproportionate impact, during this time period.

Importantly, these high impact vaccine-skeptical articles were coming from mainstream sources. Even though a much greater fraction of content from low-credibility domains (66%) was estimated to meaningfully reduce vaccination intentions compared to content from high-credibility domains (21%), high-credibility domains posted a larger total number of articles, and those articles received far more views on average. As a result, low-credibility domains were only responsible for 9.3% of the total negative impact on vaccination intentions. 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, were the sources of URLS that had the biggest negative impact on vaccination intentions (see SI Figure S9 for full list of most harmful domains).

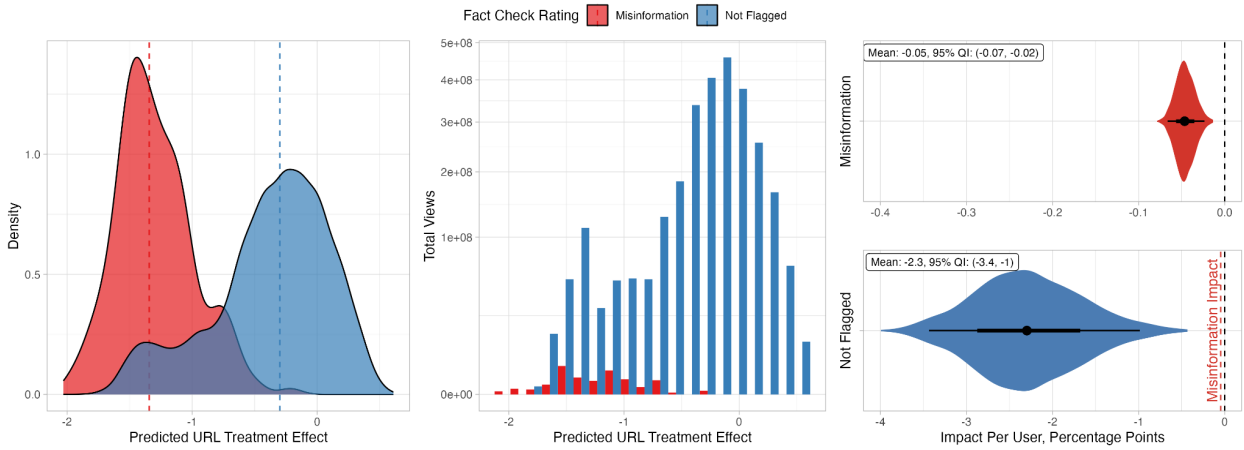

Figure 4: A, B) Distribution of predicted URL treatment effect on vaccination intentions for misinformation (shown in red) vs. vaccine-skeptical content not flagged by fact-checkers (shown in blue). Panel A shows the density plots for predicted treatment effects. Dashed lines represent the medians of 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. C) Impact for fact-checked misinformation vs. content not-flagged by fact-checkers. We subset to content predicted to be harmful (i.e. below the midpoint of the predicted Crowdsourced Aggregate Score; see S4.4 for exploration of other cutoffs). Shown is the distribution of total impact across all misinformation and vaccine-skeptical URLs, respectively, normalized by the number of US Facebook users. The point-estimates (in black) 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.

Discussion

Here we shed new light on long-standing questions about the causal effect of social media on large-scale societal outcomes. We estimate that harmful content on Facebook lowered US vaccination intentions by 2.3 percentage points per Facebook user. However, contrary to popular wisdom, we show

this effect was driven predominantly by vaccine-skeptical content from mainstream sites, rather than outright misinformation 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 approaches 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.[2,40–45] Even automated systems designed to detect and limit the spread of fake news online primarily use databases of fact-checked claims as training data.[25,46–48] Given that we find misinformation *was* substantially harmful conditional on exposure, any such efforts taken by Facebook likely improved vaccination rates to some degree. That said, we cannot tell to what extent such interventions were responsible for the low levels of exposure to misinformation, versus misinformation simply having low reach organically (as found, e.g., with political misinformation in 2016, which was prior to the widespread use of anti-misinformation tactics).[9,49] More importantly, none of these veracity-oriented interventions are likely to have reduced the spread of the type of content identified as most negatively impactful in our analyses: factually accurate but nonetheless vaccine-skeptical stories published by mainstream outlets including the Pulitzer-prize winning *Chicago Tribune*. Had exposure to this content been prevented, we estimate that vaccination rates would have been 2.3% higher on average among Facebook's 233 million US users - translating into approximately 5 million more vaccinated Americans. Assuming that 248 additional vaccinations translate into an additional life saved as estimated in Barro (2022),[50] this implies that approximately 21,000 lives could have been saved if this content had not been published or allowed to spread unchecked.[51]

Thus, our results highlight the need to consider the reach and harmfulness of content, in addition to its 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. Thus, researchers and technology companies should move beyond a narrow focus on veracity, and devote more attention to understanding, tracking - and potentially intervening on - harmful content irrespective of veracity. Similarly, 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 most people only read the headlines and context is lost. Rather than focusing exclusively on the accuracy of the facts they report, journalists should also consider whether the resulting stories will leave readers with an accurate worldview.

Of course, when considering efforts to reduce the reach of content that is harmful but not outright misinformation, it is essential to balance the desire to reduce harm against the importance of free expression. Deciding how to weigh these competing values is an extremely challenging normative question with no straightforward solution. However, an informed discussion of this tradeoff is impossible without being able to quantify harm.

To that end, the framework that we introduce here provides, for the first time, a tractable approach for quantifying harm at scale: crowdsourcing and natural-language processing, calibrated to ground-truth treatment effects from randomized survey experiments. Although, as past research has shown, laypeople often overstate the magnitude and direction of individual treatment effects,[52,53] here we show that the layperson ratings are able to predict variation across treatment effects with high accuracy. These judgments can then be used to predict treatment effects for new samples of content, and NLP methods

make it possible to extend these predictions to an enormous scale of posts. Although future work is needed to assess the extent to which this crowd prediction approach generalizes to topics beyond COVID vaccination, we are optimistic that this method could provide a way to predict the impact of content at scale for a variety of other potential harms, such as political polarization or support for undemocratic practices.[54–56]

From a theoretical perspective, our approach demonstrates how it is possible to discover which content causes harm from the "bottom-up," rather than relying on the (potentially biased) inclinations of researchers or technology company employees. In doing so, we address the "stimulus-as-fixed" fallacy, a common threat to generalizability in social science research, and contribute to the growing literature on doing causal inference using latent-dimensions of treatments in large scale social data.[57,58] 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 a much larger set of stimuli. We also contribute 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.[4,38,47,59–61]

Of course, our work also has important limitations. First, our survey 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 our survey experiments were run in mid 2022. Ideally, the experiments would happen in real time (e.g. if our approach was applied by technology companies). To help address concerns regarding the delay in the present data, we perform several robustness checks which re-examine our results with contemporaneous data for experimental effects and engagement (as a proxy for exposure), respectively, which show similar patterns (see S5.1-2). Second, our work measures survey intentions to take a COVID-19 vaccine, rather than actual vaccination behavior, and thus, could be overstated. However, Athey et. al (2023) found that survey and behavioral measures of COVID-19 vaccination are substantially correlated and in particular, found that a 1pp increase in vaccination intentions measured via Facebook surveys corresponded to a 0.6pp increase in actual county-level vaccine takeup rate.[62] Similarly, meta-analysis shows a .55 ratio of between intentions and behavior across a variety of interventions.[63] Multiplying our results by this 0.6 scaling factor would suggest that Facebook content lowered vaccination intentions by 1.38pp rather than 2.3pp, translating to 13,000 lives saved. Third, our Facebook data included only link content and did not contain information about native video, photo, or text-only content. Thus, our overall finding is a lower bound of the total amount of vaccine-skeptical and misinformation content on Facebook. It is however possible that misinformation (compared to factual information) was relatively more prevalent among non-link-based content. Future research should examine whether non-link content about vaccines showed different patterns than the ones found in our analysis.

Another potential limitation is that while our experimental participants were randomly exposed to content in a survey context, vaccine-hesitant users on Facebook might have actively sought out anti-vaccine content or been selectively targeted to see it by Facebook's algorithm. Although we do not find evidence that treatment effects (conditional on exposure) differed significantly based on participant characteristics (including pre-treatment vaccine attitudes), we cannot rule out the possibility that *exposure* to anti-vaccine content was concentrated in users who were likely going to refuse the vaccine anyway. To investigate this issue, in SI Section S4.7 we analyze the extent to which exposure to harmful content was concentrated among different demographic populations. As one might expect, very conservative users had information diets composed of the greatest proportion of content predicted to be harmful (27%). However, all political groups saw at least 10% of such content and - perhaps most concerningly - 23% of content

viewed by apolitical Facebook users (who make up appr 75% of the total user base) was predicted to be harmful. Furthermore, the fact that over 20% of Facebook's US population viewed the *Chicago Tribune* "healthy doctor dies" story suggests that vaccine-skeptical content achieved broad popularity in at least some cases. Nonetheless, understanding how repeated exposure to misinformation and vaccine-skeptical messages might change cumulative impact is a key direction for future research.

In this work, we present the first attempt to quantify the harm caused by misinformation and vaccine-skeptical content on Facebook during the first quarter of 2021. This analysis required a number of assumptions: for example regarding how mid 2022 treatment effects translate into early 2021 treatment effects, the extent of exposure implied by each Facebook view, how effect sizes on intentions translate into effect sizes on behavior, whether exposure changes downstream information-seeking (especially off-platform), and the effect of repeated exposure. Naturally, making different assumptions could lead to different conclusions. We see our primary contribution as the introduction of a *framework* for evaluating harm, more so than the specific numerical estimate that we arrive at regarding Facebook in early 2021. Our framework allows future researchers and policymakers to evaluate harm under any set of assumptions they find reasonable, and determine how these changes affect the conclusions. In doing so, we allow policy makers to make decisions about mitigating harm that are based on evidence and quantitative assessments, rather than simply on intuition.

Methods

I. Facebook Exposure Data

IA. 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.[31] 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 (i) descriptive characteristics of URLs, like the title, description, domain, and third-party fact-checker ratings of the URLs and (ii) 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 2023.

Facebook also added differentially private noise to the engagement-related columns of the dataset.[64] 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 et al 2021 for a reference for the calculation of these confidence intervals and deeper discussion of the URL Shares dataset)[33].

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).[32] However, since our analysis is largely concerned only with the

top-viewed stories, this bias does not come into play. However, the *overall* impact estimates (e.g. our estimate of how much Facebook lowered anti-vaccine content per Facebook user) are likely conservative, since anti-vaccine URL content that was not highly shared are not included in our analysis.

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 1) posted for the first time between 1/1/21 and 3/31/21, 2) had "vaccin* or "vax" in their title, description, or URL, and 3) were primarily popular in the US. 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.[31]

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.[35,65]

IB. 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 S4.1.

IC. Low Credibility Domains (Exposure)

We use a list of 2170 domains from Lasser et al (2023) labeled as "unreliable."[66] 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 (Krippendorf's alpha = .84), but unlike NewsGuard, the lists are available publicly and for free.[67] The full list is available on OSF: https://osf.io/68mn9/?view_only=42e73620bf4e4caeb9502f90d9742bdf.

II. 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. 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/?view_only=42e73620bf4e4caeb9502f90d9742bdf).

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, but were not required to, play the video. 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,[68] 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 S1.3-4).

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 S1.3). 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 S1.4.2).

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 clips from cable TV. 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/?view_only=42e73620bf4e4caeb9502f90d9742bdf.

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),[37] and our list of low-quality domains was adapted from the Iffy index (which is a subset of the low-quality domain list from Lasser et al (2021) which we use to classify our Facebook URLs.)[66] The full list of domains can be found on our OSF site https://osf.io/68mn9/?view_only=42e73620bf4e4caeb9502f90d9742bdf. We then filtered out URLs that were irrelevant, redundant, or out-of-date, leaving us with 191 candidate URLs. We randomly sampled the remaining URLs, stratified on domain type and topic, to gather a final list of 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 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: 1) surprising, 2) plausible, 3) favorable to Democrats (vs. Republicans), and 4) 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/?view_only=42e73620bf4e4caeb9502f90d9742bdf.

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 "potentially misleading" by both fact-checkers, 19 were fact-checked misleading by one fact-checker, and 49 were fact-checked as true. We labeled the URL "misinformation" if it had been fact-checked as 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).

IID. 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.[69] See S1.1.1 for exact wording.

IIE. 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 S1.2.1.

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 Hewitt et al (2023), which examined heterogeneity in political ad treatment effects.[70] 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.[71] The formal model is specified in S1.2.2.

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 S1.5.

III. 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 5. For more detail see the following sections.

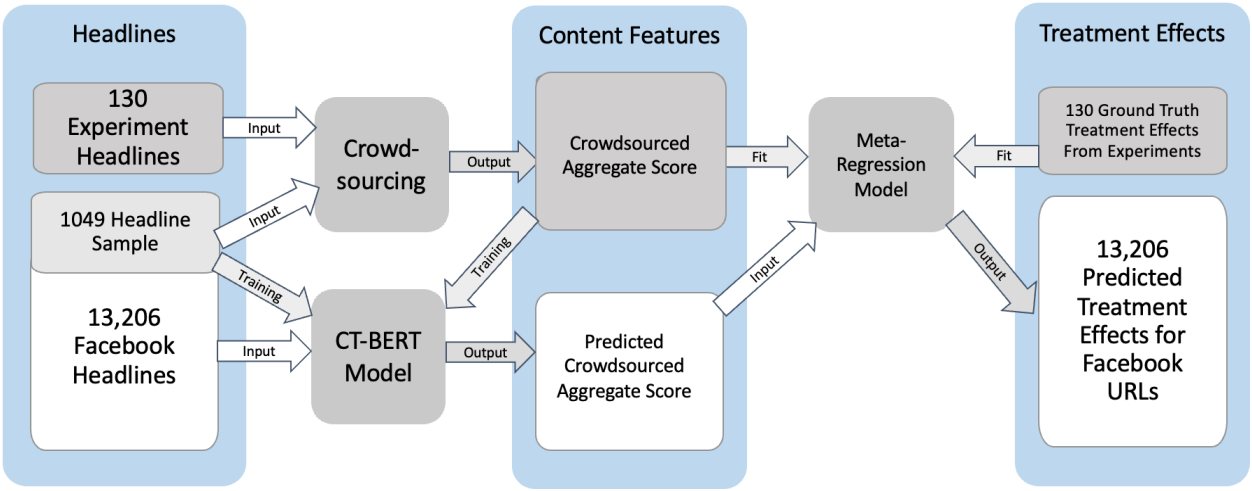

Figure 5: Methodological pipeline for predicting treatment effects for our 13,206 Facebook URLs from our 130 experimental treatment effects using crowdsourcing and machine learning.

IIIA. Crowdsourcing

We solicited crowdsourced judgments from lay people recruited from CloudResearch's Amazon Mechanical Turk platform about the predicted persuasive impact of each of the 130 vaccine-related items from Studies 1 and 2. In addition, we collected their judgments of the harmful-to-health rating and accuracy rating for each item (for precise wording, see S2.1). 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 the single "less vs. more willing to vaccinate" question alone (see S2.3 for comparison).

IIIB. 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 IIE, in which we predicted the treatment effect using content-level features, except we use crowdsourced judgments instead of content-level descriptive features. The formal model is presented in S2.2.

IIIC. 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.[72]

Annotation Procedure

We randomly sampled 1200 of the 13,206 Facebook URLs for labeling, upweighting URLs that were 1) more highly viewed, 2) from different clusters, and 3) from clusters that contained more fact-checked false content. We removed duplicates and near duplicates of the same headline, leaving us with 1139 URLs.

We then used Amazon Mechanical Turk's CloudResearch Platform to label the i) less more vax, ii) harmful-to-health, and iii) accuracy ratings of each URL using the same questions as described in Methods Section IIIA. 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 S3.3 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.[39] 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.[73]

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 1) and a binary classification model predicting whether URL's Crowdsourced Aggregate Score is below the scale midpoint of 3 (content which we deem as "harmful", i.e. likely to lower vaccination intentions, shown in Table 2). An analysis of the performance of the model alternative cutoffs for the aggregate score can be found in S3.6. For reference, the aggregate score model is measured on a 1-5 scale.

	rMSE	**Correlation**	**MAE**	**Accuracy (within .5 of true value)**	**Accuracy (within 1 of true value)**
Crowdsourced Aggregate Score	0.33	0.87	0.25	86%	99%

Table 1: Performance for a model predicting the crowdsourced aggregate score of the Facebook URL

	Accuracy	**AUROC**	**False Positive Rate**	**True Positive Rate**
Binary	90.1%	97%	4%	80%

Classification Task: Harmful vs. Not				

Table 2: Performance for a model predicting the binary harmful vs. not harmful rating of the Facebook URL

IV. 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 IIIB 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.

VI. 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).[74]

Since content that substantially lowered vaccination intentions is the focus of our analysis, we filter to Facebook URLs that we classify as "harmful," 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, 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. Analysis of the full population of URLs and results for different thresholds can be found in S.4.4.

References

1. Lazer, D. M. J. *et al.* The science of fake news. *Science* **359**, 1094–1096 (2018).
2. Shao, C., Ciampaglia, G. L., Flammini, A. & Menczer, F. Hoaxy: A platform for tracking online misinformation. in *Proceedings of the 25th international conference companion on world wide web* 745–750 (2016).
3. Shao, C. *et al.* Anatomy of an online misinformation network. *Plos One* **13**, e0196087 (2018).
4. Tschiatschek, S., Singla, A., Gomez Rodriguez, M., Merchant, A. & Krause, A. Fake News Detection in Social Networks via Crowd Signals. in *Companion Proceedings of the The Web Conference 2018* 517–524 (International World Wide Web Conferences Steering Committee, 2018).
5. Vosoughi, S., Roy, D. & Aral, S. The spread of true and false news online. *Science* **359**, 1146–1151 (2018).
6. Wang, Y., McKee, M., Torbica, A. & Stuckler, D. Systematic literature review on the spread of health-related misinformation on social media. *Soc. Sci. Med.* **240**, 112552 (2019).
7. Allen, J., Howland, B., Mobius, M., Rothschild, D. & Watts, D. J. Evaluating the fake news problem at the scale of the information ecosystem. *Sci Adv* **6**, eaay3539 (2020).
8. Altay, S., Nielsen, R. K. & Fletcher, R. Quantifying the "infodemic": People turned to trustworthy news outlets during the 2020 coronavirus pandemic. *J. Quant. Descr. Digit. Media* **2**, (2022).
9. Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B. & Lazer, D. Fake news on Twitter during the 2016 U.S. presidential election. *Science* **363**, 374–378 (2019).
10. Guess, A., Nagler, J. & Tucker, J. Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Sci Adv* **5**, eaau4586 (2019).
11. Pennycook, G. & Rand, D. G. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* **188**, 39–50 (2019).
12. Pennycook, G. & Rand, D. G. The psychology of fake news. *Trends Cogn. Sci.* **25**, 388–402 (2021).
13. Van Der Linden, S., Maibach, E., Cook, J., Leiserowitz, A. & Lewandowsky, S. Inoculating against misinformation. *Science* **358**, 1141–1142 (2017).
14. Van der Linden, S. *et al.* How can psychological science help counter the spread of fake news? *Span. J. Psychol.* **24**, e25 (2021).
15. Altay, S., Berriche, M. & Acerbi, A. Misinformation on misinformation: Conceptual and methodological challenges. *Soc. Media Soc.* **9**, 20563051221150412 (2023).
16. Puri, N., Coomes, E. A., Haghbayan, H. & Gunaratne, K. Social media and vaccine hesitancy: new updates for the era of COVID-19 and globalized infectious diseases. *Hum. Vaccines Immunother.* **16**, 2586–2593 (2020).
17. Aw, J., Seng, J. J. B., Seah, S. S. Y. & Low, L. L. COVID-19 Vaccine Hesitancy—A Scoping Review of Literature in High-Income Countries. *Vaccines* **9**, 900 (2021).
18. Bridgman, A. *et al.* The causes and consequences of COVID-19 misperceptions: Understanding the role of news and social media. *Harv. Kennedy Sch. Misinformation Rev.* **1**, (2020).
19. Linden, S. van der. We need a gold standard for randomised control trials studying misinformation and vaccine hesitancy on social media. *BMJ* **381**, p1007 (2023).
20. Loomba, S., de Figueiredo, A., Piatek, S. J., de Graaf, K. & Larson, H. J. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nat Hum Behav* (2021).
21. de Saint Laurent, C., Murphy, G., Hegarty, K. & Greene, C. M. Measuring the effects of misinformation exposure and beliefs on behavioural intentions: A COVID-19 vaccination study. *Cogn. Res. Princ. Implic.* **7**, 87 (2022).
22. Guess, A. M., Nyhan, B., O'Keeffe, Z. & Reifler, J. The sources and correlates of exposure to vaccine-related (mis)information online. *Vaccine* **38**, 7799–7805 (2020).

23. Judd, C. M., Westfall, J. & Kenny, D. A. Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem. *J. Pers. Soc. Psychol.* **103**, 54 (2012).
24. Yarkoni, T. The generalizability crisis. *Behav. Brain Sci.* **45**, e1 (2022).
25. Patwa, P. *et al.* 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* 21–29 (Springer, 2021).
26. Parks, M. Few facts, millions of clicks: Fearmongering vaccine stories go viral online. *NPR March* **25**, (2021).
27. Borges do Nascimento, I. J. *et al.* Infodemics and health misinformation: a systematic review of reviews. *Bull. World Health Organ.* **100**, 544–561 (2022).
28. Chen, E., Jiang, J., Chang, H.-C. H., Muric, G. & Ferrara, E. Charting the Information and Misinformation Landscape to Characterize Misinfodemics on Social Media: COVID-19 Infodemiology Study at a Planetary Scale. *JMIR Infodemiology* **2**, e32378 (2022).
29. Cinelli, M. *et al.* The COVID-19 social media infodemic. *Sci. Rep.* **10**, 1–10 (2020).
30. Pierri, F. *et al.* One Year of COVID-19 Vaccine Misinformation on Twitter: Longitudinal Study. *J. Med. Internet Res.* **25**, e42227 (2023).
31. Messing, S. *et al. Facebook Privacy-Protected Full URLs Data Set*. (2020).
32. Allen, J., Mobius, M., Rothschild, D. M. & Watts, D. J. Research note: Examining potential bias in large-scale censored data. *Harv. Kennedy Sch. Misinformation Rev.* (2021).
33. Guess, A., Aslett, K., Tucker, J., Bonneau, R. & Nagler, J. Cracking open the news feed: Exploring what us Facebook users see and share with large-scale platform data. *J. Quant. Descr. Digit. Media* **1**, (2021).
34. Vincent, E. M., Théro, H. & Shabayek, S. Measuring the effect of Facebook's downranking interventions against groups and websites that repeatedly share misinformation. *Harv. Kennedy Sch. Misinformation Rev.* (2022) doi:10.37016/mr-2020-100.
35. Facebook. *Facebook's Third-Party Fact-Checking Program*.
36. Allen, J., Arechar, A. A., Pennycook, G. & Rand, D. G. Scaling up fact-checking using the wisdom of crowds. *Sci. Adv.* **7**, eabf4393 (2021).
37. Pennycook, G. & Rand, D. G. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proc Natl Acad Sci U A* **116**, 2521–2526 (2019).
38. Kim, J., Tabibian, B., Oh, A., Schölkopf, B. & Gomez-Rodriguez, M. *Leveraging the Crowd to Detect and Reduce the Spread of Fake News and Misinformation*. (2018).
39. Müller, M., Salathé, M. & Kummervold, P. E. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *ArXiv Prepr. ArXiv200507503* (2020).
40. Aslett, K., Guess, A. M., Bonneau, R., Nagler, J. & Tucker, J. A. News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions. *Sci. Adv.* **8**, eabl3844 (2022).
41. Guess, A. M. *et al.* A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proc. Natl. Acad. Sci.* **117**, 15536–15545 (2020).
42. Clayton, K. *et al.* Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Polit. Behav.* **42**, 1073–1095 (2020).
43. Oeldorf-Hirsch, A., Schmierbach, M., Appelman, A. & Boyle, M. P. The Ineffectiveness of Fact-Checking Labels on News Memes and Articles. *Mass Commun. Soc.* **23**, 682–704 (2020).
44. NewsGuard - Combating Misinformation with Trust Ratings for News. *NewsGuard* https://www.newsguardtech.com/.
45. Hartwig, K., Doell, F. & Reuter, C. The Landscape of User-centered Misinformation Interventions -- A Systematic Literature Review. Preprint at https://doi.org/10.48550/arXiv.2301.06517 (2023).
46. Ruchansky, N., Seo, S. & Liu, Y. CSI: A Hybrid Deep Model for Fake News Detection. in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* 797–806

(Association for Computing Machinery, 2017).
47. Shabani, S. & Sokhn, M. Hybrid Machine-Crowd Approach for Fake News Detection. in *2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC)* 299–306 (2018).
48. Kazemi, A., Garimella, K., Gaffney, D. & Hale, S. A. Claim matching beyond English to scale global fact-checking. *ArXiv Prepr. ArXiv210600853* (2021).
49. Allcott, H. & Gentzkow, M. Social Media and Fake News in the 2016 Election. *J Econ Perspect* **31**, 211–236 (2017).
50. Barro, R. J. Vaccination rates and COVID outcomes across US states. *Econ. Hum. Biol.* **47**, 101201 (2022).
51. Lewis, A. M., Tanya. How to Compare COVID Deaths for Vaccinated and Unvaccinated People. *Scientific American* https://www.scientificamerican.com/article/how-to-compare-covid-deaths-for-vaccinated-and-unvaccinated-people/.
52. Graham, M. H. & Coppock, A. Asking About Attitude Change. *Public Opin. Q.* **85**, 28–53 (2021).
53. Milkman, K. L. *et al.* A 680,000-person megastudy of nudges to encourage vaccination in pharmacies. *Proc. Natl. Acad. Sci.* **119**, e2115126119 (2022).
54. Timberg, C., Dwoskin, E. & Albergotti, R. Inside Facebook, Jan. 6 violence fueled anger, regret over missed warning signs. *Washington Post* https://www.washingtonpost.com/technology/2021/10/22/jan-6-capitol-riot-facebook/ (2021).
55. Bavel, J. J. V., Rathje, S., Harris, E., Robertson, C. & Sternisko, A. How social media shapes polarization. *Trends Cogn. Sci.* **25**, 913–916 (2021).
56. Kubin, E. & von Sikorski, C. The role of (social) media in political polarization: a systematic review. *Ann. Int. Commun. Assoc.* **45**, 188–206 (2021).
57. Fong, C. & Grimmer, J. Causal inference with latent treatments. *Am. J. Polit. Sci.* (2021).
58. Egami, N., Fong, C. J., Grimmer, J., Roberts, M. E. & Stewart, B. M. How to make causal inferences using texts. *Sci. Adv.* **8**, eabg2652 (2022).
59. Cheng, J. & Bernstein, M. S. Flock: Hybrid Crowd-Machine Learning Classifiers. in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* 600–611 (Association for Computing Machinery, 2015). doi:10.1145/2675133.2675214.
60. Groh, M., Epstein, Z., Firestone, C. & Picard, R. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proc. Natl. Acad. Sci.* **119**, e2110013119 (2022).
61. Srivastava, M., Hashimoto, T. & Liang, P. Robustness to Spurious Correlations via Human Annotations. in *Proceedings of the 37th International Conference on Machine Learning* 9109–9119 (PMLR, 2020).
62. Athey, S., Grabarz, K., Luca, M. & Wernerfelt, N. Digital public health interventions at scale: The impact of social media advertising on beliefs and outcomes related to COVID vaccines. *Proc. Natl. Acad. Sci.* **120**, e2208110120 (2023).
63. Webb, T. L. & Sheeran, P. Does changing behavioral intentions engender behavior change? A meta-analysis of the experimental evidence. *Psychol. Bull.* **132**, 249–268 (2006).
64. Dwork, C. Differential privacy. in *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33* 1–12 (Springer, 2006).
65. About Fact-Checking on Facebook and Instagram | Meta Business Help Center. https://www.facebook.com/business/help/2593586717571940?id=673052479947730.
66. Lasser, J. *et al.* Social media sharing of low-quality news sources by political elites. *PNAS Nexus* **1**, pgac186 (2022).
67. Lin, H. *et al.* High level of agreement across different news domain quality ratings. (2022).
68. Aronow, P. M., Kalla, J., Orr, L. & Ternovski, J. Evidence of rising rates of inattentiveness on Lucid in 2020. (2020).
69. Broockman, D. E., Kalla, J. L. & Sekhon, J. S. The design of field experiments with survey outcomes: A framework for selecting more efficient, robust, and ethical designs. *Polit Anal* **25**,

435–464 (2017).
70. Hewitt, L. *et al.* How experiments help campaigns persuade voters: evidence from a large archive of campaigns' own experiments. *Am. J. Polit. Sci.* **Forthcoming**, (2023).
71. Borenstein, M., Hedges, L. V., Higgins, J. P. & Rothstein, H. R. *Introduction to meta-analysis*. (John Wiley & Sons, 2021).
72. Sagi, O. & Rokach, L. Ensemble learning: A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **8**, e1249 (2018).
73. Hossain, T. *et al.* 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).
74. U.S.: Facebook users 2018-2027. *Statista* https://www.statista.com/statistics/408971/number-of-us-facebook-users/.

Supplementary Information for "Quantifying the Impact of Misinformation and Vaccine-Skeptical Content on Facebook"

Contents

S1 Experiments **S2**

S1.1 Study Variable Definitions . . . S2

S1.1.1 Outcome Variable . . . S2

S1.1.2 Pre-Treatment Covariates . . . S3

S1.2 Model . . . S3

S1.2.1 Study 1: Effect of Misinformation . . . S3

S1.2.2 Studies 1 and 2: Stimulus-Level Heterogeneity . . . S4

S1.3 Balance Checks . . . S4

S1.4 Differential Attrition . . . S5

S1.4.1 Study 1 Attrition . . . S5

S1.4.2 Study 2 Attrition . . . S6

S1.5 Additional Survey Results . . . S6

S1.5.1 Study 1, Additional Outcome Variables . . . S6

S1.5.2 Disaggregated moderators . . . S7

S1.5.3 Study 2, Additional moderators . . . S7

S1.5.4 Subject Level Heterogeneity . . . S9

S1.6 Study 1 Headlines . . . S10

S1.7 Study 2 Headlines . . . S12

S2 Crowdsourcing **S20**

S2.1 Variable Definition . . . S20

S2.2 Formal Model . . . S20

S2.3 Crowdsourcing Performance . . . S21

S2.4 Unadjusted Correlation . . . S21

S3 NLP Model **S22**

S3.1 Performance . . . S22

S3.2 Predicting Component Variable Performance . . . S23

S3.3 Alternative Models . . . S23

S3.4 Predicted URL Impact S24
S3.5 Predicted Treatment Effect CIs S24
S3.6 Cutoff Tuning S24

S4 Facebook Impact **S26**

S4.1 Top Viewed URLs S26
S4.2 Impact Calculation S27
S4.3 Quantile Intervals for Impact Estimates S27
S4.4 Threshold for Harmful URLs S28
S4.5 Impact by Domain Type S28
S4.6 Most Harmful Domains S30
S4.7 Subject-Level Heterogeneity S30
S4.8 Most Harmful Stories S31
S4.9 Model, Lives Saved S32

S5 Robustness **S34**

S5.1 Contemporaneous Engagement Estimates S34
S5.2 Contemporaneous Treatment Effect Estimates S34

S1 Experiments

This section provides a deeper description of our survey experiments, as well as supplementary results. We pre-registered an analysis plan for both experiments (see `https://osf.io/68mn9/?view_only=42e73620bf4e4caeb9502f90d9742bd`). However, due to space and readability constraints, we only included a subset of our pre-registered analyses in the main text. No pre-registered analysis that was not included changes the interpretation of our results. We report these as well as other supplementary results in Section S1.5.

S1.1 Study Variable Definitions

S1.1.1 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?”

S1.1.2 Pre-Treatment Covariates

We fit our models using the following pre-treatment covariates in order to increase statistical power [9].

- 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

S1.2 Model

S1.2.1 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.

$$\begin{aligned} Y_i = \beta_0 + \beta_1 \text{pre\_vax\_index}_i + \alpha \text{vax\_treat}_i + \beta_2 \text{pre\_vax\_index}_i \times \text{vax\_treat}_i \\ + \beta_3 \times \text{gender}_i + \beta_4 \times \text{age}_i + \beta_5 \times \text{pol}_i + \epsilon_i \\ \alpha = \alpha_0 + \alpha_k \\ \alpha_k \sim N(0, \sigma) \end{aligned}$$

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.

S1.2.2 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:

$$\begin{aligned} Y_{ij} = \beta_{0j} + \sum_{k \in K} \theta_{jk} \text{treat}_{ijk} + \beta_{1j} \text{pre\_vax\_index}_{ij} + \beta_{3j} \times \text{gender}_{ij} \\ + \beta_{4j} \times \text{age}_{ij} + \beta_{5j} \times \text{pol}_{ij} + \epsilon_{ij} \end{aligned}$$

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.

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

- j indexes study
- k indexes stimulus
- μ 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.

S1.3 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.

Table S1: 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 S2: 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

S1.4 Differential Attrition

Again, we test for differential attrition with a model predicting whether or not a person attrited (1 = attritted, 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.

S1.4.1 Study 1 Attrition

The overall attrition rate is 2.9%. 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.

Table S3: 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.

S1.4.2 Study 2 Attrition

We test for differential attrition with a model predicting whether or not a person attrited (1 = attritted, 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%. 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 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 (β=-.59,p =.007).

Table S4: Attrition Check, Study 2

Variable	Coefficient	p.value	p.adj
Treatment (vs Control)	-0.03	<0.001	<0.001
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.

S1.5 Additional Survey Results

Here, we report additional results from our survey experiments. Most of these results stem from our original pre-registration, except where they are marked as "Post-Hoc."

S1.5.1 Study 1, Additional Outcome Variables

For Study 1, we estimated the main effect of vaccine misinformation on vaccination intentions. The main outcome that we pre-registered is our "Vaccine Intent Index", composed of the average of 4 questions gauging participants willigness to take a vaccine (see S1.1.1 for details). We report the disaggregated variables composing the index here. Note that fewer participants were eligible for the "Vaccine Intent (First Dose)" and "Booster Intent" questions than the other two, since those questions were answered only by participants who had not yet received a first dose or booster, respectively, and thus they are relatively less powered.

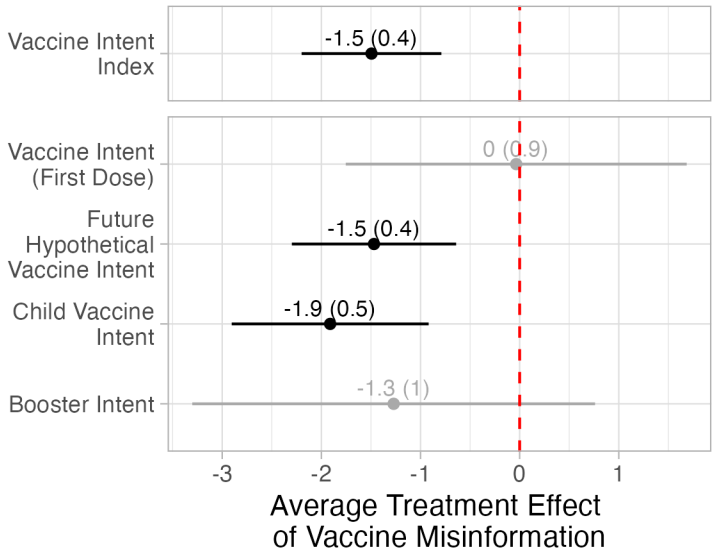

Figure S1: Average Treatment Effect of Vaccine Misinformation on each question in our our vaccine-index, respectively.

S1.5.2 Disaggregated moderators

To assess what features of vaccine-related content best explained variation in the treatment effects, we performed a random-effects meta-regression with our vaccine-index as the outcome variable and content-level features as moderators (see Section S1.2.2 for model specification.)

In the main text, we pooled Study 1 and Study 2 data for maximum power. Here, we report the disaggregated moderators for each study separately, as well as the pooled effect. We pre-registered these specific moderators and the analysis plan for Study 2; we apply the same analysis procedure for Study 1 and our pooled analyses, but they were done post-hoc.

S1.5.3 Study 2, Additional moderators

Surprising X Harmful-to-Health As can be seen in Figure S3, in Study 1, we found that the extent to which an item was i) surprising and ii) harmful-to-health predicted variation in the treatment effect. Based on this finding, we pre-registered a model for Study 2 that contained both harmful to health, surprising, and their interaction as moderators. The results are summarized in Table S5. Only harmful-to-health was significant.

Table S5: Study 2: Surprising x Harmful to Health Model

Term	Coefficient	SE	p.value
Surprising	0.09	0.11	0.45
Harmful-to-Health	-0.30	0.12	0.01
Surprising:Harmful-to-Health	0.06	0.15	0.70

Site-Quality Interactions Additionally, we pre-registered an exploratory model that included an interaction between site quality and all covariates. Because we only vetted these stimuli post-hoc, in our pre-registration we believed that site quality would be a proxy for misinformation (although later analysis shows that it was a poor approximation). No terms were significant, although one should note that the model is significantly underpowered. Nonetheless, we report it for transparency.

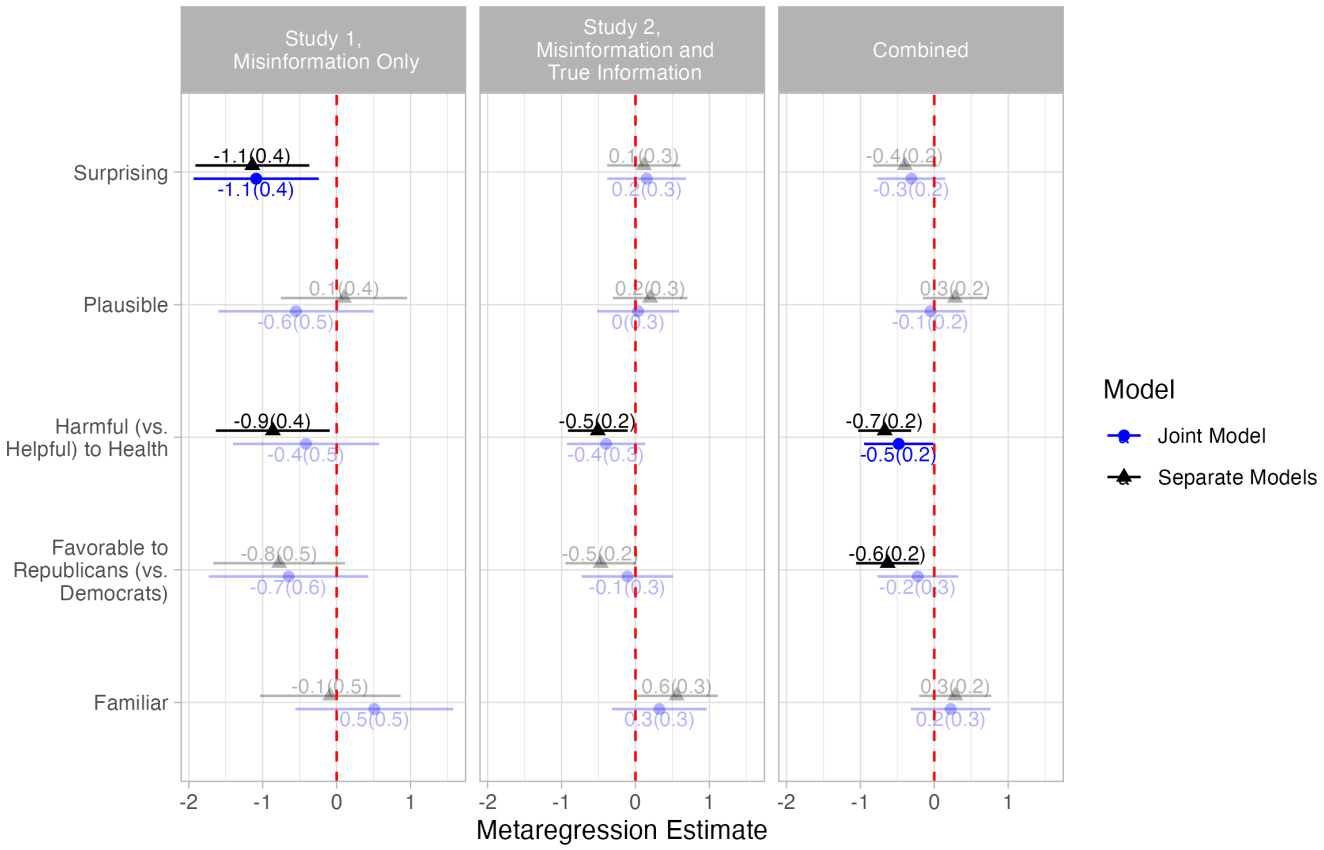

Figure S2: The coefficients from respective meta-regressions testing how different features of content moderate the treatment effects in i) Study 1, ii) Study 2, and a iii) multi-level meta-regression combining both studies. The black points show the results from a model testing all moderators separately, and the blue points are from a joint model that contains all 5 moderators. Coefficients with $p < .05$ are bolded.

Table S6: Study 2: Domain Type x Covariates

Term	Coefficient	SE	p.value
familiar	0.14	0.15	0.34
site_quality	-0.03	0.13	0.82
demrep	-0.09	0.17	0.61
harmful_to_health	-0.22	0.17	0.20
surprising	0.01	0.13	0.92
plausible	0.02	0.14	0.86
familiar:site_quality	0.20	0.15	0.17
site_quality:demrep	-0.19	0.17	0.25
site_quality:harmful_to_health	-0.10	0.17	0.57
site_quality:surprising	0.03	0.13	0.85
site_quality:plausible	-0.11	0.14	0.45

Extra Moderators Finally, post-hoc, we collected additional labels from Lucid to assess whether dimensions of content that we missed explained variation in the effects. These additional moderators are shown in Figure S3, along with the logged engagement (engagementL) that the URL received on Facebook, collected form CrowdTangle. The only additional variables that explained the effect were harm-related, providing additional support for our main results.

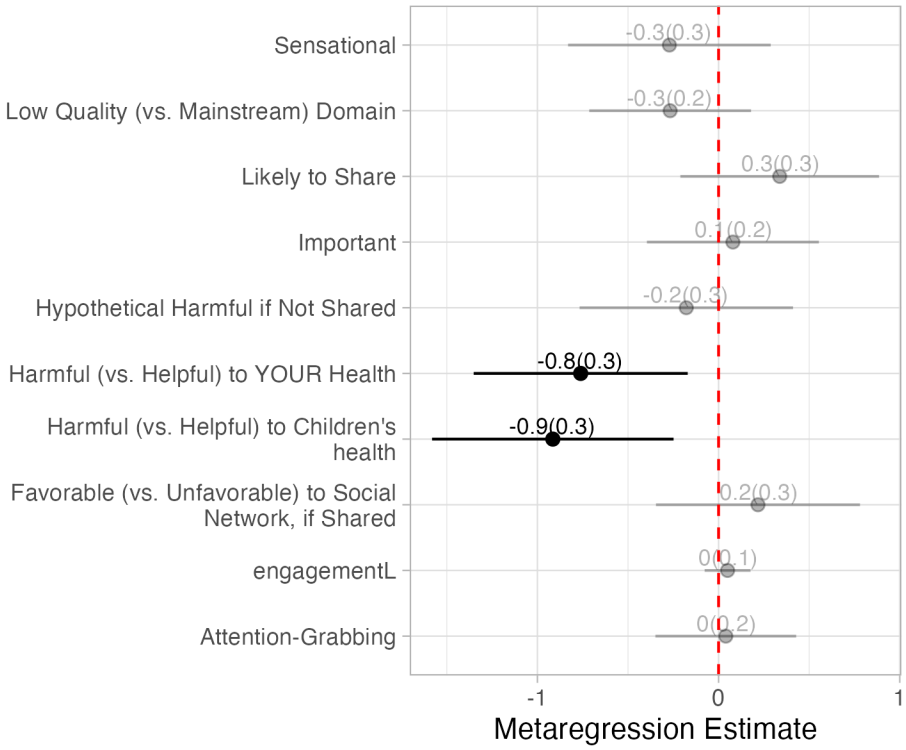

Figure S3: The coefficients from respective meta-regressions testing how different features of content moderate the treatment effects in Study 2. Coefficients with p < .05 are bolded.

S1.5.4 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.

We find no strong evidence of individual-level heterogeneity. Although both gender and unvaccinated status are almost statistically significant (p=.07), after adjusting for multiple comparisons, the evidence for substantial heterogeneity is unconvincing. 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 [8].

Table S7: Study 1: Individual Level Heterogeneity

study	Variable	Estimate	p.value	p.adj
Study 1	Is Male	1.23	0.07	0.20
Study 1	Age (Standardized)	0.28	0.40	0.48
Study 1	Is Democrat (vs. Republican)	-0.67	0.32	0.48
Study 1	Is Unvaccinated	1.40	0.07	0.20
Study 1	Pre-Treatment Vaccination Intentions (Standardized)	-0.34	0.31	0.48
Study 1	Is Boosted	-0.12	0.85	0.85

We repeat the same heterogeneity analysis for Study 2. 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 S8: Study 2: Individual Level Heterogeneity

study	Variable	Estimate	p.value	p.adj
Study 2	Is Male	1.60	0.11	0.61
Study 2	Age (Standardized)	0.16	0.75	0.90
Study 2	Is Democrat (vs. Republican)	-0.72	0.47	0.71
Study 2	Is Unvaccinated	1.15	0.31	0.61
Study 2	Pre-Treatment Vaccination Intentions (Standardized)	-0.58	0.24	0.61
Study 2	Is Boosted	-0.09	0.93	0.93

S1.6 Study 1 Headlines

Table S9: Experiment 1 Headlines

ID	Headline
exp1_id_1	Data Suggests Vaccine-Jabbed Individuals Are More Likely to Catch Omicron
exp1_id_10	mRNA inventor says young adults shouldn't have to get COVID vaccine
exp1_id_11	Doctor Warns Stillbirths Are Rampant Among Fully Vaccinated Mothers, Launches Investigation
exp1_id_12	Tweet: Vaccines and Omicron have the same symptoms
exp1_id_13	CHD to Sue FDA for 'Recklessly Endangering' Children if Agency Authorizes Pfizer Vaccine for Children 5 to 11 Years Old
exp1_id_14	Pfizer adds ingredient used to stabilize heart attack victims in vax for kids
exp1_id_15	Aaron Rodgers: "If the vaccine is so great, then how come people are still getting COVID and spreading COVID and, unfortunately, dying from COVID?"
exp1_id_16	CA Governor Gavin Newsom MIA, Rumors He's Suffering from Vaccine Injury From October Booster Shot
exp1_id_17	Japan drops vax rollout, goes to Ivermectin, ENDS COVID almost overnight
exp1_id_18	FB post: most vaccine deaths are among vaccinated
exp1_id_19	Plane crashes linked to the Covid vaccine jab
exp1_id_2	Tucker Carlson: How many Americans have died after taking the COVID vaccine?
exp1_id_20	BREAKING: Air Traffic Controllers In Jacksonville, FL Staged A Walkout Yesterday In Response To The Vaccine Mandate
exp1_id_21	Washington Wizards Star Bradley Beal Questions vaccine efficacy
exp1_id_22	Candace Owens: the Taliban is banning the Covid Vaccine
exp1_id_23	New Israeli Study Finds Fully Vaccinated People are at "Greater Risk of Hospitalization" and 13 TIMES MORE LIKELY to Catch Covid-19 Than Those Who Have Recovered and Have Natural Immunity

exp1_id_24	FDA ‘playing bait and switch’ with Americans, tricking them into believing shots currently being offered have been granted full approval when they have not
exp1_id_25	Insane hypocrisy: Biden White House staff not required to be vaxxed
exp1_id_26	Pfizer CEO had to cancel a planned trip to Israel because he was not vaccinated
exp1_id_27	There is an 82% miscarriage rate among women who got the vaccine between 30 days and 20 weeks pregnant.
exp1_id_28	Marjorie Tayor Greene: Biden & Dems are coming to your door and forcing you to take the vax
exp1_id_29	Liz Wheeler: "This peer-reviewed scientific study shows that COVID-19 vaccine causes two deaths for every three lives it saves"
exp1_id_3	Tucker Carlson: Maybe [the COVID-19 vaccine] doesn’t work, and they’re simply not telling you that
exp1_id_30	Bill Filed In Washington Would Authorize ‘Strike Force’ To ‘Involuntarily Detain’ Unvaccinated Families
exp1_id_31	SCIENCE vs POLITICS: Dr. Geert Vanden Bossche On Why The COVID Vaccine Is A Bad Idea For Kids
exp1_id_32	Govt. Data Reveals Over 946K People Suffered Injuries Or Death From COVID-19 Vaccinations
exp1_id_33	Official UK Government Data Suggests Fully Vaxxinated Brits Will Develop Acquired Immunodeficiency Syndrome (AIDS) by the End of February 2022
exp1_id_34	The COVID-19 vaccines’ “spike protein is very dangerous, it’s cytotoxic.”
exp1_id_35	Johnson & Johnson: ’Kids Shouldn’t Get A F*cking [COVID] Vaccine;’ There are "Unknown Repercussions"
exp1_id_36	The Lancet Publishes Medical Prof’s Warning That Natural Immunity Has Made Vaccine Mandates Irrelevant
exp1_id_37	43 Studies on Vaccine Efficacy that Raise Doubts on Vaccine Mandates
exp1_id_38	Fox host says “the risk to children from the vaccine outweighs the risk to COVID”
exp1_id_39	People in Their Twenties Have 7 Times Higher Risk of Death After Vaccination Than From COVID’ - RAIR
exp1_id_4	CDC Data Show COVID-19 Mass Vaccination Has Had No Measurable Impact on COVID-19 Mortality in the U.S.
exp1_id_40	Nicki Minaj: Vaccine causes erectile dysfunction
exp1_id_5	Report Shows Nearly 300 Athletes Worldwide Collapsed or Suffered Cardiac Arrests after Taking COVID Vaccine This Year – Many Died
exp1_id_6	MRNA Technology Inventor Dr. Robert Malone Warns Parents: Vaccines Could Permanently Damage Children’s Critical Organs.
exp1_id_7	Johns Hopkins Doctor: "I have never seen a vaccine that forced me to wear a mask and maintain my social distance, even when you are fully vaccinated"
exp1_id_8	“I’m Over COVID, What the F**k Is the Use of Boosters?” – Bill Maher Goes Off on Medical Establishment, Experimental Vaccines in Latest Interview

exp1_id_9	4th COVID Booster Shot Could Cause ‘Immune System Fatigue,’ Scientists Say

S1.7 Study 2 Headlines

Table S10: Experiment 2 Headlines

ID	Headline	Description	Fact-Check Rating
exp2_id_1	Pfizer CEO paints ‘ideal’ future of Covid vaccination — RT World News	Pfizer CEO Albert Bourla has expressed hope that yearly vaccinations would be more palatable to skeptics than boosters every few months	True
exp2_id_10	A 4th dose of Covid-19 vaccine will be needed, Pfizer’s CEO says, but the company is working on a shot to handle all variants | CNN	Pfizer is also hoping to make a vaccine that will protect against all variants. "If we are able to achieve that, then I think it is very easy to follow and remember so that we can go back to really...	True
exp2_id_11	An 85-year-old man in India says he got 12 Covid vaccine shots, and still wants more.	The retired postman said he gamed the offline vaccination drive to keep boosting regularly — long before booster shots were a thing anywhere.	True
exp2_id_12	Biden will get additional Covid booster if his doctor recommends it, White House says	The White House said President Joe Biden will receive an additional Covid-19 booster shot if his doctor recommends it, after the US Food and Drug Administration expanded emergency authorization of ...	True
exp2_id_13	Starting later this week, some at-risk Americans become eligible for a 4th shot.	A change in federal recommendations allows moderately or severely immunocompromised people to receive their boosters five months after their third shot, which was a part of the primary immunization.	True
exp2_id_14	Fourth COVID vaccine still doesn’t stop Omicron, new Israeli study...	The study raised questions about Israel’s decision to be the first in the world to offer a second booster shot — and fourth overall — to its over-60 population.	True
exp2_id_15	’We can’t vaccinate the planet every six months,’ says Oxford vaccine scientist | CNN	A leading expert who helped create the Oxford-AstraZeneca Covid-19 vaccine said Tuesday that giving everyone in the world booster shots multiple times a year is not feasible.	True
exp2_id_16	Those who got J&J’s COVID vaccine should seriously consider a Pfizer or Moderna booster, experts say	For those who got the J&J COVID vaccine, experts recommend a Pfizer or Moderna booster for better protection.	True
exp2_id_17	3-year-old girl dies of heart attack one day after taking COVID vaccine - Geller Report	Forcing children to have this vaccine is a crime of unprecedented magnitude.	False/Misleading

exp2_id_18	JUST IN: Florida Department of Health to Recommend Against Covid Vax For Healthy Children	Florida's Surgeon General Dr. Joseph Ladapo will recommend against Covid vaccination for healthy children. WPTV reported: Healthy children in Florida shouldn't get the COVID-19 vaccine. That was th...	True
exp2_id_19	Kansas Senate Passes Bill to Authorize the Prescriptions of Ivermectin and Hydroxychloroquine and Child Vaccine Exempt	Kansas state senators passed a bill early Thursday that would authorize the prescriptions of off-label drugs for Covid-19 treatment, such as Ivermectin and hydroxychloroquine. The bill also exempts...	True
exp2_id_2	Moderna Asks FDA For Emergency Authorization of Second Covid Booster Shot For All Adults: NYT	Moderna asked the FDA to approve a second Covid booster for all adults, the New York Times reported. The New York Times reported: Moderna said late Thursday that it asked the Food and Drug Administ...	True
exp2_id_20	Ronald McDonald House to Evict 4-Year-Old Leukemia Patient, Parents Over Vaccine Status	As a kid, I always enjoyed visiting McDonald's – both for its food and its fundraiser for the Ronald McDonald House.	False/Misleading
exp2_id_21	MIT Scientist Warns Parents NOT TO GIVE CHILDREN Vaccine, Could Cause 'Crippling' Neurodegenerative Disease In Young People - Geller Report	'It's outrageous to give these vaccines to young people. It doesn't make any sense.' [Children] "have a very low risk of dying from Covid." For young people, the benefits derived from the vaccines ...	False/Misleading
exp2_id_22	Florida Health Department Recommends AGAINST Healthy Kids Getting The COVID Vaccine	The Florida Department of Health has officially recommended against giving COVID vaccines to healthy children, boldly breaking from federal guidelines. [. . .]	False/Misleading
exp2_id_23	CDC: 80.2% of Americans 5 and Older Have Had at Least One COVID-19 Shot	(Photo by Paul Hennessy/SOPA Images/LightRocket via Getty Images) (CNSNews.com) - As of Thursday, 80.2 percent of Americans 5 years old and older have had at least one shot of a COVID-19 vaccine an...	True
exp2_id_24	Undercover Students EXPOSE Colorado High School Vaccine Clinic For Administering Vaccinations Without Parental Consent - Even AFTER School Superintendent ASSURED This Would Not Happen	Libs of TikTok is dropping bombs once again. On Monday night, the independent journalist exposed yet another issue within US Public School system – this time, posting a damning thread that includes...	False/Misleading
exp2_id_25	Moderna says its Covid-19 vaccine performs as well in children as it does in adults	Moderna announced interim results of its Covid-19 vaccine for children younger than 6 on Wednesday.	True
exp2_id_26	CDC recommends Pfizer-BioNTech booster for 12-to-17-year-olds	Walensky says booster dose will provide 'optimized protection' against omicron variant.	True

exp2_id_27	Covid Updates: Number of Hospitalized Young Children Who've Tested Positive Is Jumping, C.D.C. Says	The Supreme Court heard arguments over President Biden's vaccine mandates, and Citigroup will dismiss unvaccinated employees by the end of the month.	True
exp2_id_28	The F.D.A. clears booster shots for 12- to 15-year-olds.	Mayor Eric Adams of New York City insisted schools would remain open despite surging Omicron cases in the city. He said remote learning has been too damaging, especially to children in low-income n...	True
exp2_id_29	COVID Vaccine in kids less effective against Omicron vs Delta, but ward off severe illness from both: study	The Pfizer/BioNTech COVID-19 vaccine provided less protection against the Omicron variant than the Delta strain in children but did protect against severe illness from both variants.	True
exp2_id_3	Pfizer CEO: Omicron Specific Covid Vaccine will be Ready by March (VIDEO)	So we're taking vaccines for a common cold with sniffles now? Pfizer CEO Albert Bourla on Monday said his company will have a vaccine specifically made for the Omicron variant by March. The top fiv...	False/Misleading
exp2_id_30	COVID worsens asthma in children; booster after infection not as beneficial vs Omicron	The following is a summary of some recent studies on COVID-19. They include research that warrants further study to corroborate the findings and that has yet to be certified by peer review.	True
exp2_id_31	Covid-19 booster raises antibody levels against Omicron for children ages 5 through 11, Pfizer and BioNTech say	A third shot of the children's dose of Pfizer/BioNTech's Covid-19 vaccine raised Omicron-fighting antibodies by 36 times in kids 5 through 11 years of age, the companies said in a news release Thur...	True
exp2_id_32	Pfizer Shot Is Far Less Effective in 5- to 11-Year-Olds Than in Older Kids, New Data Show	While protection against hospitalization is still strong, the vaccine offered almost no protection against infection, even just a month after full vaccination.	True
exp2_id_33	Father of Two Young Children Denied Heart Transplant for Being Unvaccinated	A 31-year-old father from Boston was removed from the heart transplant list because he refused to receive the experimental COVID-19 vaccine 'which puts him at high risk for adverse reactions and ev...	False/Misleading
exp2_id_34	Former Head Of UK Vaccine Taskforce Says It's Now A 'Waste of Time' To Keep Vaccinating People - News Punch	According to the former chairman of the UK's Vaccine Taskforce, it's now a "waste of time" to keep vaccinating people against covid.	False/Misleading
exp2_id_35	South Carolina Senate Passes Legislation that Bans Covid-19 Vaccine Requirements	The South Carolina Senate passed a bill on Wednesday prohibiting businesses from refusing to serve unvaccinated people and preventing government employees, first responders, and students from takin...	True

exp2_id_36	RNA Vaccine Inventor Dr. Robert Malone: 'Time of choosing' for CDC scientists after bombshell NYT report - Geller Report	Has there been a more cowardly and corrupt collective than the medical profession during this crackdown on our individual freedoms. The truckers have done the work the doctors should have done year...	False/Misleading
exp2_id_37	Twitter Users Notice Something Really Creepy in Lori Lightfoot's Latest "Get Vaxed" Photo	It's weird enough that Lori Lightfoot is using "dollar bills" in a weird vaccine stunt, but things got even more bizarre when Twitter users noticed something	True
exp2_id_38	Saskatchewan Residents No Longer Need To Show Vaccination Status To Enter Places	Saskatchewan has lifted its public health order that required residents to show proof of COVID-19 vaccination or a negative test to enter most businesses. The vaccine passport system, which was bro...	True
exp2_id_39	Zients Warns U.S. Won't Have Enough Vaccines for a 4th Dose for General Population If CDC Recommends It	A health worker applies a fourth dose of the Pfizer COVID-19 vaccine on March 22, 2022 in San Salvador, El Salvador. El Salvador begun the application of a fourth dose of the COVID-19 vaccine. (Pho...	True
exp2_id_4	Covaxin Booster Shown To Neutralise Both Omicron And Delta Variants Of COVID-19: Bharat Biotech	NA	False/Misleading
exp2_id_40	Military Doctor's Testimony: Ordered by High-Level Command to Keep Quiet on Vaccine Issues	Uh oh...	False/Misleading
exp2_id_41	The U.S. is extending a vaccine rule for international travelers at its land borders.	Unlike air travelers entering the United States, land and ferry travelers will still not have to show a recent negative coronavirus test to cross the border.	True
exp2_id_42	Germany posts a one-day record in cases even as it plans to lift restrictions.	NA	True
exp2_id_43	LA County moves to shift who can discipline unvaccainted workers as sheriff refuses to enforce the mandate	The authority to impose penalties for noncompliance with Los Angeles County's vaccine mandate for public workers soon could shift way from department heads who may not be carrying it out – namely ...	True
exp2_id_44	DC won't require COVID vaccination proof for entertainers but will impose restriction for their patrons	Washington D.C. will not require entertainers to show a proof of a coronavirus vaccination when performing at venues the district.	True
exp2_id_45	US Navy discharges 240 service members for refusing Covid-19 vaccine | CNN Politics	The US Navy said Wednesday that it has discharged 240 service members for refusing to get the Covid-19 vaccine as required by the Pentagon's vaccine mandate.	True
exp2_id_46	A Florida public health official is put on leave after emailing his staff to urge vaccination.	Dr. Raul Pino, the state public health administrator in Orlando, told employees that without a good reason, not being vaccinated against the coronavirus was irresponsible.	True

exp2_id_47	White House tells agencies to delay vaccine mandate after court win	The White House told federal agencies on Thursday to hold off on reinstating a coronavirus vaccination mandate for millions of employees, hours after an appeals court rejected an earlier injunction...	True
exp2_id_48	A D.C. bar that was closed after defying vaccination requirements has a new life as a conservative rallying point.	Senator Rand Paul and other Republicans are lionizing The Big Board, a family-owned bar a mile from the U.S. Capitol.	True
exp2_id_49	Poorer nations reject millions of expiring Covid vaccine doses – UN — RT World News	The program to help poorer nations to vaccinate their populations against Covid-19 is facing a problem, as many donations have a remaining shelf life too short to be properly distributed, a UN offi...	True
exp2_id_5	[VIDEO] Trump: Politicians Who Refuse to Say If They've Received Vaccine Booster Are 'Gutless'	Former President Donald Trump has said that politicians who refuse to declare whether or not they have had booster shots when questioned in interviews are	True
exp2_id_50	Dr. Malone Reveals That a Top Investor of Spotify Also Top Investor of Moderna (VIDEO)	Dr. Robert Malone, the inventor of mRNA vaccine technology, joined with Tucker Carlson Today to discuss cancel culture following leftists call to cancel Joe Rogan about his stand on the COVID vacci...	True
exp2_id_51	Ep. 1731 Stunning Vaccine Info Emerges - The Dan Bongino Show	For show notes, visit https://bongino.com/ep-1731-stunning-vaccine-info-emerges Check out our Clips channel for video highlights https://rumble.com/c/DanBonginoShowClips Sign up to receive Dan's daily	False/Misleading
exp2_id_52	British Medical Journal Demands Immediate Release of All COVID-19 Vaccine, Treatment Data	...regulators are not there to "dance to the tune of rich global corporations and enrich them further" but to protect the general public's health...	True
exp2_id_53	LA Times Columnist Says Mocking Anti-Vaxxers' Deaths Is "Necessary"	The first incarnation was called 'Why Shouldn't We Dance On The Graves of Anti-Vaxxers?'	True
exp2_id_54	We Have to Speak in Code — Citizens With Reported Vaccine Injuries Being Silenced on Social Media	People who report vaccine injuries are reportedly being silenced across multiple social media platforms.	False/Misleading
exp2_id_55	BREAKING: FDA Executive Officer Makes SHOCKING Admission About Covid Vaccine In Secretly Recorded Video [WATCH]	Yesterday, shocking audio was released by Project Veritas that exposed the FDA for lying to the American people about the [. . .]	False/Misleading
exp2_id_56	Tipping Point - Daniel Horowitz - The Military's Vaccine Injury Cover-Up	Daniel Horowitz - The Military's Vaccine Injury Cover-Up	False/Misleading
exp2_id_57	Israel Considers 4th Vaccine Dose, but Some Experts Say It's Premature	Some scientists warn that too many shots might actually harm the body's ability to fight the Covid-19 virus. But Israeli experts say there isn't time to wait.	True

exp2_id_58	No, Bob Saget and Betty White's deaths were not due to the COVID-19 vaccine - Poynter	Following the shocking deaths of beloved stars Betty White and Bob Saget, false claims circulated online blaming the COVID-19 booster shot for their deaths.	True
exp2_id_59	Lisa Shaw: Presenter's death due to complications of Covid vaccine	Lisa Shaw developed headaches shortly after being vaccinated against Covid-19, an inquest hears.	True
exp2_id_6	Ex-FDA Chief: 'Heading Into the Fall, I Suspect a Lot of Americans Will Want to Get Another Vaccine'	__alt__ (CNSNews.com) - Pfizer's COVID vaccine "is really a six-month vaccine in providing meaningful protection against meaningful infection," former FDA Commissioner Scott Gottlieb told CBS's "Fa...	True
exp2_id_60	Dr. Kizzy Corbett on omicron: 'A boosted, vaccinated person' will fight this virus away	Dr. Kizzmekia Corbett and her NIH team developed the Moderna vaccine two years ago, before the first case of Covid-19 was even confirmed in the U.S. Now a Harvard University School of Public Health...	True
exp2_id_61	Opinion | People are wrecking their bodies trying to 'detox' from the Covid vaccine	Research suggests vaccinations may have averted up to 140,000 deaths in the United States.	True
exp2_id_62	Israel says 500K have received 4th vaccine dose	Israel's Health Ministry says more than 500,000 people have received a 4th vaccine dose	True
exp2_id_63	Pope Francis calls anti-vaccine sentiment 'baseless' in his annual state-of-the-world speech	Vaccines aren't a "magical means of healing," the pope said, but still represent the best way of fighting the virus.	True
exp2_id_64	Man in Germany gets 90 COVID-19 shots to sell forged passes	A 60-year-old man allegedly had himself vaccinated against COVID-19 dozens of times in Germany in order to sell forged vaccination cards with real vaccine batch numbers to people not wanting to get...	True
exp2_id_65	The culprit in COVID and vaccine harm: Micro blood clots	Joel S. Hirschhorn shares many studies showing how spike proteins 'screw up' capillaries	False/Misleading
exp2_id_66	Major Insurance Company Estimates 2.5-3 Million People in Germany 'Under Treatment For Side Effects of Vaccination After Covid Shot' - Geller Report	When historians write this terrible, horrible story, it will be viewed as the largest medical experiment in human history that failed on a monumental scale.	False/Misleading
exp2_id_67	34-Yr-Old Mom and Loving Wife Dies of Rare Brain Bleed One Day After Getting COVID Jab...Coroner Blames COVID Vaccine	#TheirLivesMattered In April 2021, Reuters wrote about a senior official for the European Medicines Agency (EMA) who admitted that there [. . .]	True
exp2_id_68	STAGGERING: 833 Athlete Cardiac Arrests and Serious Issues, 540 Dead, Following Covid Injection - Geller Report	The medical establishment, now an arm of the Biden regime thanks to Obamacare, remains not just silent, but complicit.	False/Misleading
exp2_id_69	Pfizer Unable to Secure COVID Vaccine Authorization in India, China Due to Side Effect Concerns - Geller Report	Drug regulators bar 2.8 billion people from company's COVID vaccine but Democrats here in America are forcing it on our children.	False/Misleading

exp2_id_7	Kathy Griffin Posts Her "4th Booster Shot" Photo And Instantly Regrets It LOL	One of the strangest things that happened with this whole COVID mess, was how obsessed leftists got over a vaccine that didn't do what it was supposed to do,	False/Misleading
exp2_id_70	WATCH: Pilot Got the Vaccine And Now He is No Longer Able to fly. - Geller Report	Pilots to sue the Biden administration.	False/Misleading
exp2_id_71	DOD/Fauci Caught In Massive Scandal Hiding Death And Damage From COVID Vaccines	WTF!!! Share my videos on all platforms, not just our hugboxes...make new accounts on their platforms and expect to be deleted. Just get it out there, we need to break through the echo chamber...we...	False/Misleading
exp2_id_72	EU Announces Investigation Into Reported Menstrual Disorders Following COVID Vaccine	This is a good thing...	True
exp2_id_73	CDC weighs increasing time between vaccine doses to lower risk of heart inflammation	U.S. health officials are considering new changes to vaccine guidance that would lengthen the amount of time between doses in order to lower the risk of heart inflammation for immunocompromised peo...	True
exp2_id_74	Covid: Woman died from rare vaccine side-effect	A coroner says Kim Lockwood, who died 10 days after receiving her Covid jab, was "extremely unlucky".	True
exp2_id_75	New Zealand Says Man's Death May Be Linked to Pfizer Vaccine	Health officials in New Zealand believe a 26-year-old man's death may be linked to Pfizer's COVID-19 vaccine.	True
exp2_id_76	Inquest finds man, 96, died after allergic reaction to Covid vaccine	Peter Jackson died after collapsing in the car park after receiving his coronavirus vaccine	True
exp2_id_77	'Preventable tragedy': Fox News silent after guest dies of Covid	Chris Hayes: For ratings, for fame, for cynical, monetary purposes, that network—overseen by CEO Suzanne Scott—has decided to fan the flames of vaccine resistance. And it's getting thousands of peo...	True
exp2_id_78	Covid-19 Vaccines Were Deadly in Rare Cases. Governments Are Now Weighing Compensation.	The U.S. and U.K. are in the very early stages of applying existing vaccine-injury programs to hundreds of claims of injury alleged from Covid-19 shots.	True
exp2_id_79	His wife died from Johnson & Johnson Covid vaccine complications. Why he's still pro-vaccine.	For the first time, Stan Thomas shares the story of how his wife, Monica Melkonian, died from vaccine-induced immune thrombotic thrombocytopenia, a rare side effect .	True
exp2_id_8	FRIGHTENING: WHO Joins EU and Changes Direction – Suddenly Warns Against Taking Continued COVID Booster Shots	On Tuesday European regulators warned that the COVID booster shots could adversely affect the immune system. This was a huge admission for European officials after pushing booster shots just weeks ...	False/Misleading

exp2_id_80	Study links Covid-19 vaccination to small increase in menstrual cycle length, but experts say it's no cause for concern	After getting a dose of Covid-19 vaccine, women had an average menstrual cycle length of about one day longer than usual, according to a study published Thursday.	True
exp2_id_81	'How do you even trust them?': Aaron Rodgers fires back at Biden after president orders QB to 'get vaxxed'	Last December, Joe Biden made a comment to a Packer's fan, "Tell that quarterback he's gotta get the vaccine." The QB took offense.	True
exp2_id_82	Jab Or Journalism?!	Allison Royal was fired from her reporter position for refusing the vaccine. Now, she's doing her own reporting in her own way. I spoke to her about Florida's 'Don't Say Gay' bill, gender and sex in e	True
exp2_id_83	Judge rules FDA can't keep vaccine docs secret 'until 2096' — RT World News	The US Food and Drug Administration (FDA) has been ordered to hasten the publication of documents related to Pfizer's coronavirus vaccine by more than one-hundred-fold, after the agency claimed the...	True
exp2_id_84	Time between Pfizer and Moderna Covid-19 vaccines can be up to 8 weeks for some people, updated CDC guidance says	The interval between first and second doses of the Pfizer and Moderna Covid-19 vaccines may be as long as eight weeks for certain people, the US Centers for Disease Control and Prevention said in v...	True
exp2_id_85	A universal vaccine could be the future of the coronavirus fight	Scientists are working to develop a "pan-coronavirus" vaccine – one that offers protection against multiple variants of the coronavirus that causes Covid-19.	True
exp2_id_86	US joins club of nations offering second booster	The US has extended the fourth jab to adults aged 50 and over. Not everyone is behind the decision.	True
exp2_id_87	CDC: Pregnant Women Are Dropping Likes Flies - News Punch	The CDC has confirmed that large numbers of pregnant woman are now "dying of Covid" and have urged them to increase their uptake of the vaccine.	False/Misleading
exp2_id_88	Paradigm Shift: RFK Jr Goes Off on Vaccine Safety Live on TV, No One Attacks Him	Robert F. Kennedy Jr. discusses the importance of vaccine safety on Tucker Carlson's show and he was actually taken seriously.	False/Misleading
exp2_id_89	FDA authorizes coronavirus vaccine boosters for 12-to-15-year-olds as schools reopen amid omicron surge	The agency also cleared booster shots for children 5 to 11 with compromised immune systems.	True
exp2_id_9	Pfizer asks the F.D.A. to authorize a second booster shot for older Americans.	The move could reignite a tortuous debate that raged last year among scientists over how often people need to bolster the vaccines' protection.	True
exp2_id_90	COVID-19 poll shows majority of parents reluctant to vaccinate kids under 5	Local experts said the poll by the Kaiser Family Foundation shows that health providers need to communicate with parents of young children so they can weigh the potential harm caused by a case of C...	True

S2 Crowdsourcing

S2.1 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 Harmful to 5 - Very Helpful). We reverse this score such 5 = Very Harmful, and 1 = Very Helpful, for consistency with the other 2 variables.
- **Accuracy** Adapted from [2]. 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)
  - 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)

S2.2 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 S1.2.2, 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 S1.2.2.

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

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

- j indexes study
- k indexes stimulus
- μ 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_s and the pseudo-R^2 (estimated below) and I^2 of the model.

S2.3 Crowdsourcing Performance

In TableS11, 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 S2.1), and iii) the model with an aggregrate crowdsourced variable as a moderator (the "Crowdsourced Aggregate Score", see above).

Note that 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 in the model not attributable to sampling variation ($\sigma_1^2 + \sigma_2^2$) [10].

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

.

Both the single and multiple question crowdsourcing models show large improvement over the baseline random effects model (ΔAIC = 11.08 and 13.65, respectively, well above the cutoff of 2 established in [7]). The multiple-question model also shows better fit (ΔAIC = 2.18), and substantially higher I^2 and pseudo R^2 values). Because of this improved fit, we use the multiple-question “Crowdsourced Aggregate Score" in for our main analysis.

	Baseline	Single Question	Multiple Questions
Intercept	-0.82	-2.67^{***}	-3.13^{***}
	(0.55)	(0.62)	(0.64)
Less vs. More Likely to Vaccinate		0.67^{***}	
		(0.19)	
Crowdsourced Aggregate Score			0.85^{***}
			(0.22)
Pseudo-R^2	NA	0.62	0.74
I^2	37.6%	18.51%	13.44%
τ^2	0.72	0.27	0.18
DF Resid.	129	128	128
AIC	430.39	418.92	416.74

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

Table S11: Model Comparison, Crowdsourcing

S2.4 Unadjusted Correlation

Additionally, we plot the correlation between the “Crowdsourced Aggregate Score" and the raw treatment effect.

It is important to note that the correlation-coefficient is attenuated because it does not appropriately correct for sampling variance in the treatment effects, unlike the meta-regression model. However, we report it for full transparency.

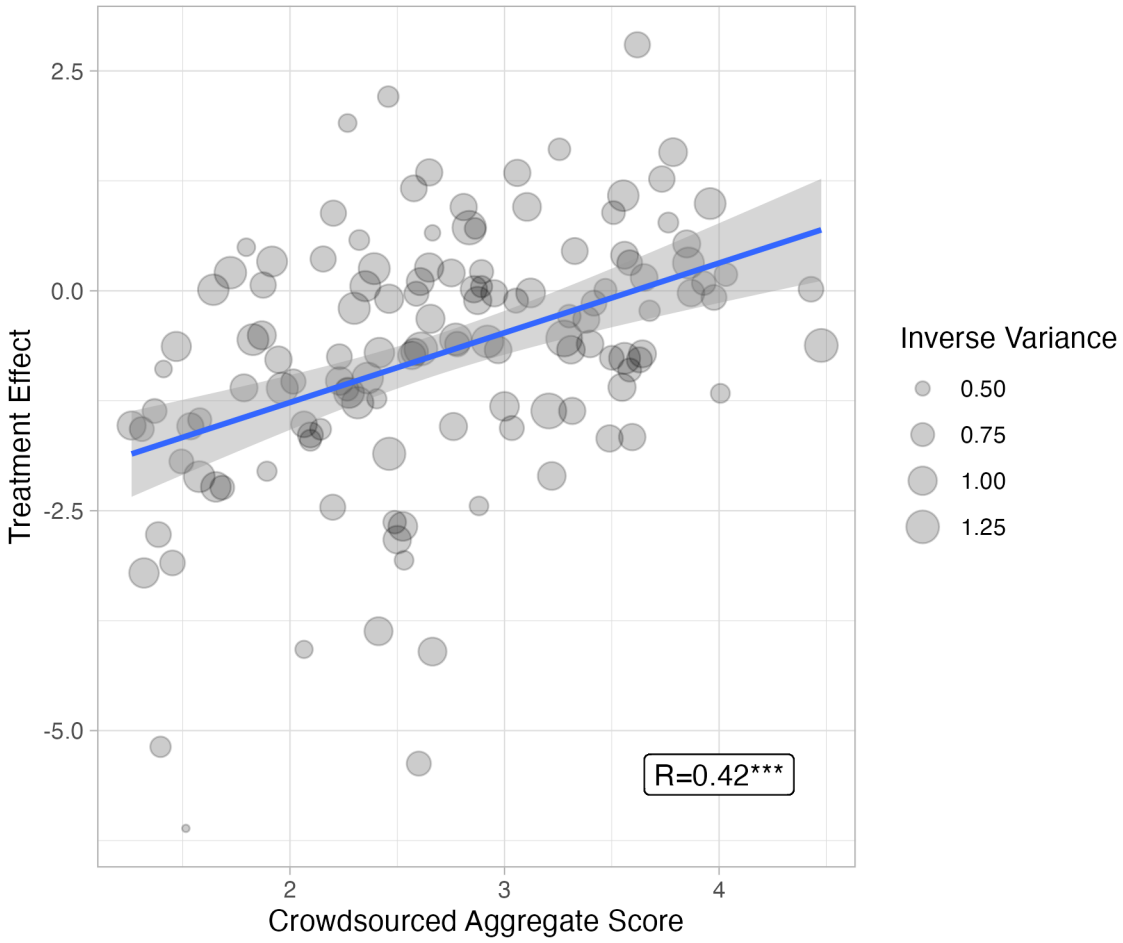

Figure S4: The correlation between the Crowdsourced Aggregate Score and the treatment effects (unadjusted for sampling variation).

S3 NLP Model

S3.1 Performance

In Table S12 We report the performance statistics for our best-performing NLP model predicting our Crowdsourced Aggregate Score, which is the average of 1) Less vs. More likely to vaccinate, 2) Harmful-to-Health (Reversed), 3) Accuracy (see Section S2.1). We train three models to predict each variable separately, and then combine them together to create our "Crowdsourced Aggregate Score." These component variable models' performance is reported in Section S3.2.

Table S12: Performance Metrics for Best Model

Variable	MSE	RMSE	MAE	Accuracy (with .5)	Accuracy (with 1)
Crowdsourced Aggregate Score	0.11	0.34	0.26	0.86	0.99

In Table S17 report the results of a binary classification task, predicting whether or not the score was above or below

the scale midpoint (3).

Table S13: Performance Metrics, Binary Classication Task (1 = Score < 3, 0 = Score >= 3)

Variable	Accuracy	AUC	F1-Score	FPR	TPR
Crowdsourced Aggregate Score	0.91	0.97	0.89	0.04	0.84

S3.2 Predicting Component Variable Performance

In Table S16, we also report the performance results from predicting the component variables for our main model.

Table S14: Performance Metrics for Intermediate Variables

Variable	MSE	RMSE	MAE	Accuracy (with .5)	Accuracy (with 1)
Less vs. More Likely to Vaccinate	0.19	0.43	0.35	0.76	0.99
Crowdsourced Accuracy	0.25	0.50	0.38	0.70	0.94
Harmful-to-Health	0.18	0.43	0.33	0.74	0.98

In Table S15, we report performance for a classification task, where we threshold each variable at the scale midpoint.

Table S15: Performance Metrics for Intermediate Variables, Binary Classification Task. Each variable was split at the scale midpoint.

Variable	Accuracy	AUC	F1-Score	FPR	TPR
Less vs. More Likely to Vaccinate	0.86	0.93	0.83	0.12	0.83
Crowdsourced Accuracy	0.86	0.91	0.74	0.07	0.69
Harmful-to-Health	0.82	0.88	0.80	0.09	0.73

S3.3 Alternative Models

We also report the results for two alternate models. “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 (reported in Section S3.1) 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 S16: 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 Aggregate Score (Single)	0.12	0.35	0.27	0.86	0.99

Table S17: 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

S3.4 Predicted URL Impact

For each URL u in our set of 13,206 vaccine-related Facebook URLs, we estimate p_u, the predicted treatment effect of exposure to the URL on intentions to take a future vaccine, measured in percentage points.

Specifically, we estimate p_u using the following equation, where i) s_u is the predicted "Crowdsourced Aggregate Score" from our best NLP model described in Section S3.1 and ii) μ is the Intercept term and β_s is the coefficient on the "Crowdsourced Aggregated Score" from the model defined in Section S2.2 and estimated in Section S2.3.

$$p_u = \mu + \beta_s \times s_u$$

S3.5 Predicted Treatment Effect CIs

In order to compute confidence intervals, we also compute draws 1000 of the treatment effect p_u by parametrically bootstrapping the coefficients using $\hat{\Sigma}_{\mu,\beta_s}$ the variance-covariance matrix of our estimated coefficients.

$$\begin{bmatrix} \beta_s^* \\ \mu^* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \beta_s \\ \mu \end{bmatrix}, \hat{\Sigma}_{\mu,\beta_s} \right)$$
$$p_u^* = \mu^* + \beta_s^* \times s_u$$

This gives us a distribution of predicted effects $\{p_{u,1}^* ... p_{u,1000}^*\}$ for each URL.

S3.6 Cutoff Tuning

For each URLs for which we have ground-truth labels, we classify a URL "Harmful" if it has a ground-truth "Crowdsourced Aggregate Score" less than 3, the scale midpoint, otherwise we classify it as "Not Harmful".

Then, to assess the performance of our model on our binary classification task, we use the predicted "Crowdsourced Aggregate Score" to predict the true binary "Harmful / Not-Harmful" 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 "Harmful." Figure S5 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 (90.1%),

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-Harmful" (and potentially even promoting vaccination) being considered “Harmful" (i.e. questioning vaccination).

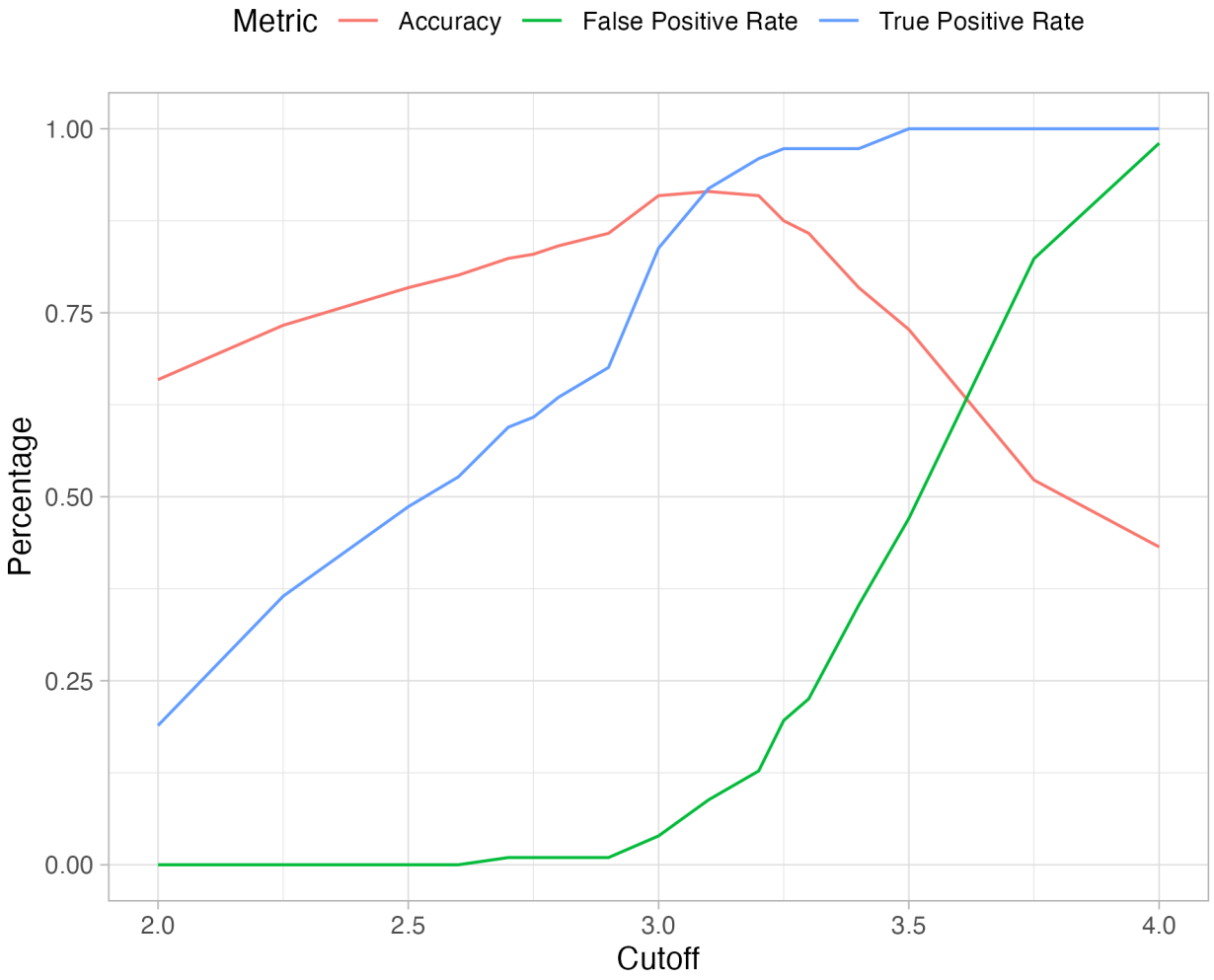

Figure S5: Performance of a model that uses the continuous predicted “Crowdsourced Aggregate Score," binarized at varying thresholds, to predict the binary “Harmful" vs “Not-Harmful" rating of a URL

S4 Facebook Impact

S4.1 Top Viewed URLs

We show here 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 (https://www.facebook.com/formedia/tools/coronavirus-resources). We show the top 10 URLs with and without these stories.

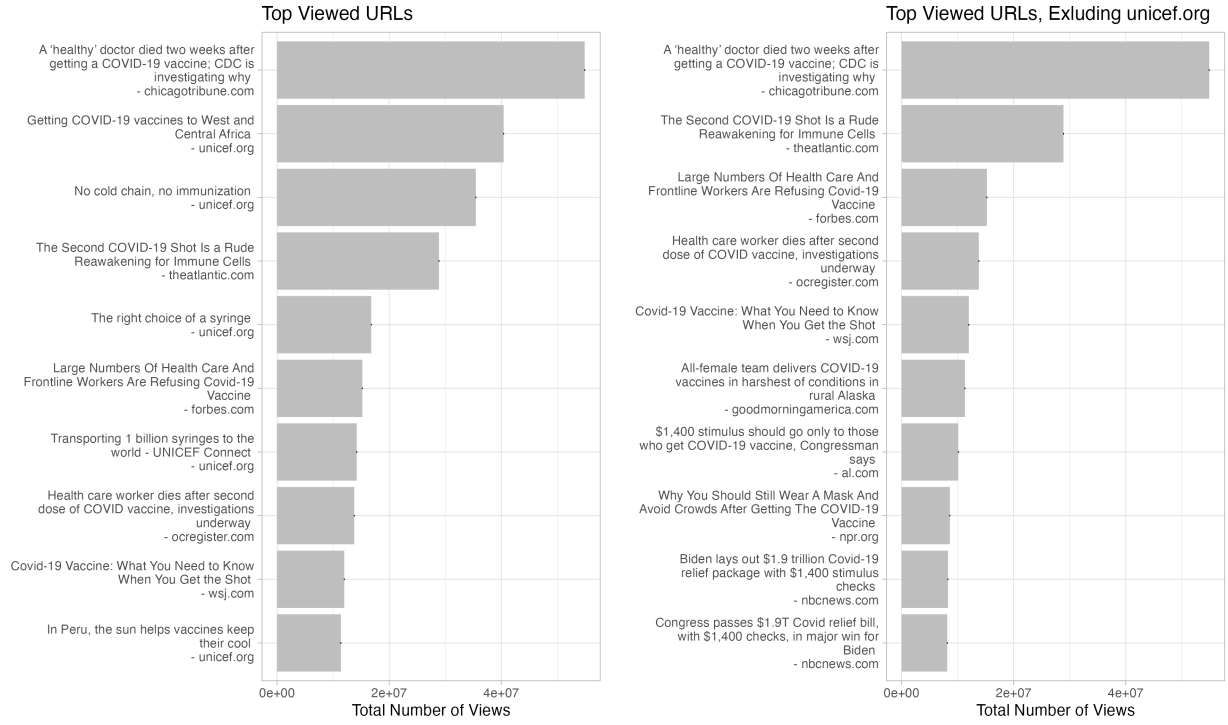

Figure S6: Top Individual URLs. Panel A shows the Top URLs including unicef.org. Panel B shows the Top URLs excluding unicef.org.

S4.2 Impact Calculation

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

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

- p_u: The predicted persuasive effect for each URL u.
- v_u: The number of views for each URL u.
- s_u: The "Crowdsourced Aggregate Score" for each URL u.
- h_u: Binary indicator for whether u is Harmful (1) or Not (0). We classify a URL as "Harmful" if $s_u < 3$ (for an exploration of other cutoffs see XX).
- m_u: Binary indicator for whether u 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 harmful and contain misinformation.
- U_{VS}: The set of all URLs designated as harmful but do not contain misinformation. We refer to this content as "Vaccine Skeptical."

Formally, this is:

$$U_M = \{u | h_u = 1 \text{ and } m_u = 1\}$$
$$U_{VS} = \{u | h_u = 1 \text{ and } m_u = 0\}$$

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

$$I_M = \frac{\sum_{u \in U_M} p_u \cdot V_u}{N_{FB}}$$

$$I_{VS} = \frac{\sum_{u \in U_{VS}} p_u \cdot V_u}{N_{FB}}$$

These equations represent the sum of the product of the predicted persuasive effect and the number of views for each harmful 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.

S4.3 Quantile Intervals for Impact Estimates

For each draw i from 1...1000 of our predicted treatment effects $p*_{u,i}$ defined in Section S3.5, we calculate overall impact I^*_M and I^*_{VS}.

From $i = 1...1000$:

$$I^*_{M,i} = \frac{\sum_{u \in U_M} p^*_{u,i} \cdot v_u}{N_{FB}}$$

$$I^*_{VS,i} = \frac{\sum_{u \in U_{VS}} p^*_{u,i} \cdot v_u}{N_{FB}}$$

We report the full distributions as well as the 95% quantile intervals for I^*_M and I^*_{VS}, respectively, in Figure 4 of the main text.

S4.4 Threshold for Harmful URLs

For our impact estimates, we subset to URLs classified as harmful, where we classify URL u as harmful if the "Crowdsourced Aggregate Score" s_u is less than 3. We choose to focus only on "harmful" URLs, rather than the whole distribution of URLs, because our research question asks how much content on Facebook lowers vaccination intentions. Furthermore, we find little evidence that any particular type of content increases vaccination intentions across our experiments. That is, items below the midpoint of "Crowdsourced Aggregate Score" significantly decreases vaccination intentions, but we do not see evidence that content above the midpoint of the scale increases vaccination intentions. Thus, we exclude this content above the scale midpoint from our overall impact estimates. By doing so, we assume that this content had an overall null effect on vaccination intentions. For models that incorporate other assumptions about this pro-vaccine content, see Figure S14.

We use 3, the scale midpoint, as a threshold for designating content as harmful in the main text. In Figure S7, we show how results would differ for different cutoffs for harmful. The left panel shows the number of URLs designated as Harmful vs. Not Harmful at different cutoffs of s_u. The right panel shows how our overall impact estimates (the sum of I_M and I_{VS}) change as the cutoff for being classified as harmful increases. As our cutoff increases, the number of URLs classified as harmful increases. However, perhaps surprisingly, although our impact estimates increase with higher thresholds (since more URLs are included and considered harmful), the width of our confidence intervals increase at higher thresholds as well. This is because our meta-regression model that uses s_u to predict p_u is more uncertain at higher values of s_u, due to less statistical power. Fewer of the 130 items we tested in our experiments was rated as likely to increase vaccination than was rated as likely to decrease vaccination intentions, and those items with high s_u had higher variance treatment effects.

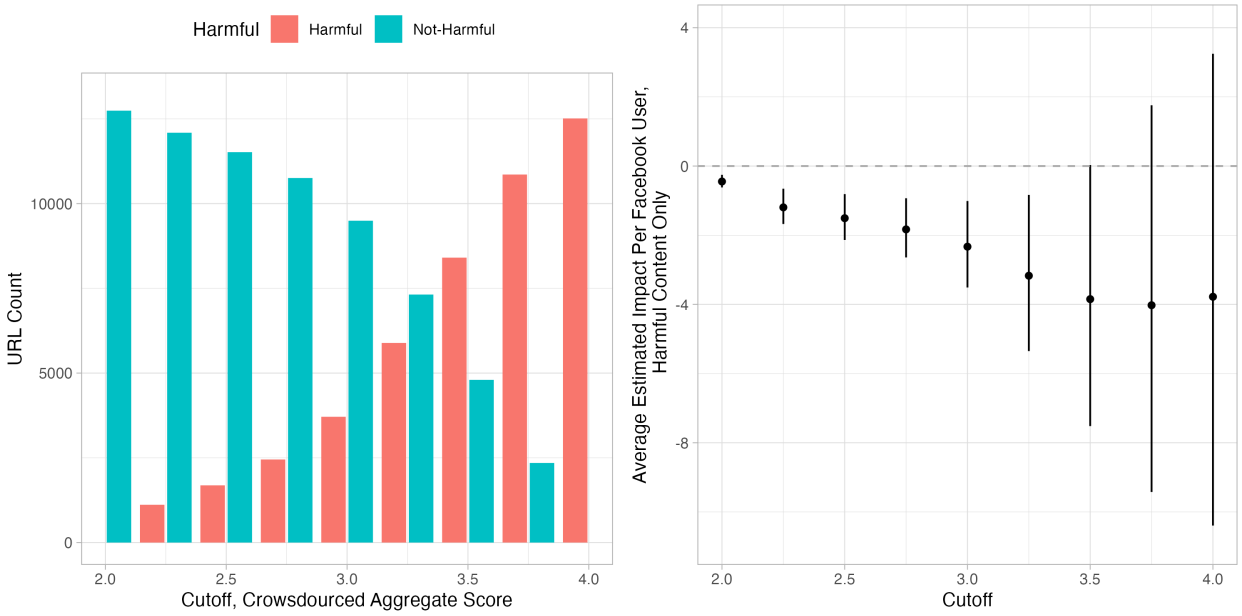

Figure S7: Varying cutoffs impact on results. **Left** How varying the cutoff affects the number of URLs considered harmful. **Right** How varying the cutoff affects overall impact estimates on vaccination intentions.

S4.5 Impact by Domain Type

In the main text, we compared the impact of outright misinformation to vaccine-skeptical content based on the rating of fact-checkers. In this section, we consider consider the relative impact of harmful vaccine content from low vs. high credibility domains. We define a domain as "Low Credibility" based on [12]; see Methods section for more details.

As can be seen in top panel of Figure S8, a larger share of the URLs published by low-credibility domains are "Harmful"

compared to the share published by high-quality domains, which mainly publish content that is neutral or the vaccine. However, because a minority of URLs are from low-credibility outlets, high credibility outlets publish an overall greater number of harmful URLs, even though most of the content they publish the vaccine. The bottom panel of Figure S8 shows the relative impact of high vs. low credibility domains for harmful vaccine content. High credibility domains have a much larger impact on overall vaccination intentions for – approximately 10X – that of low credibility domains.

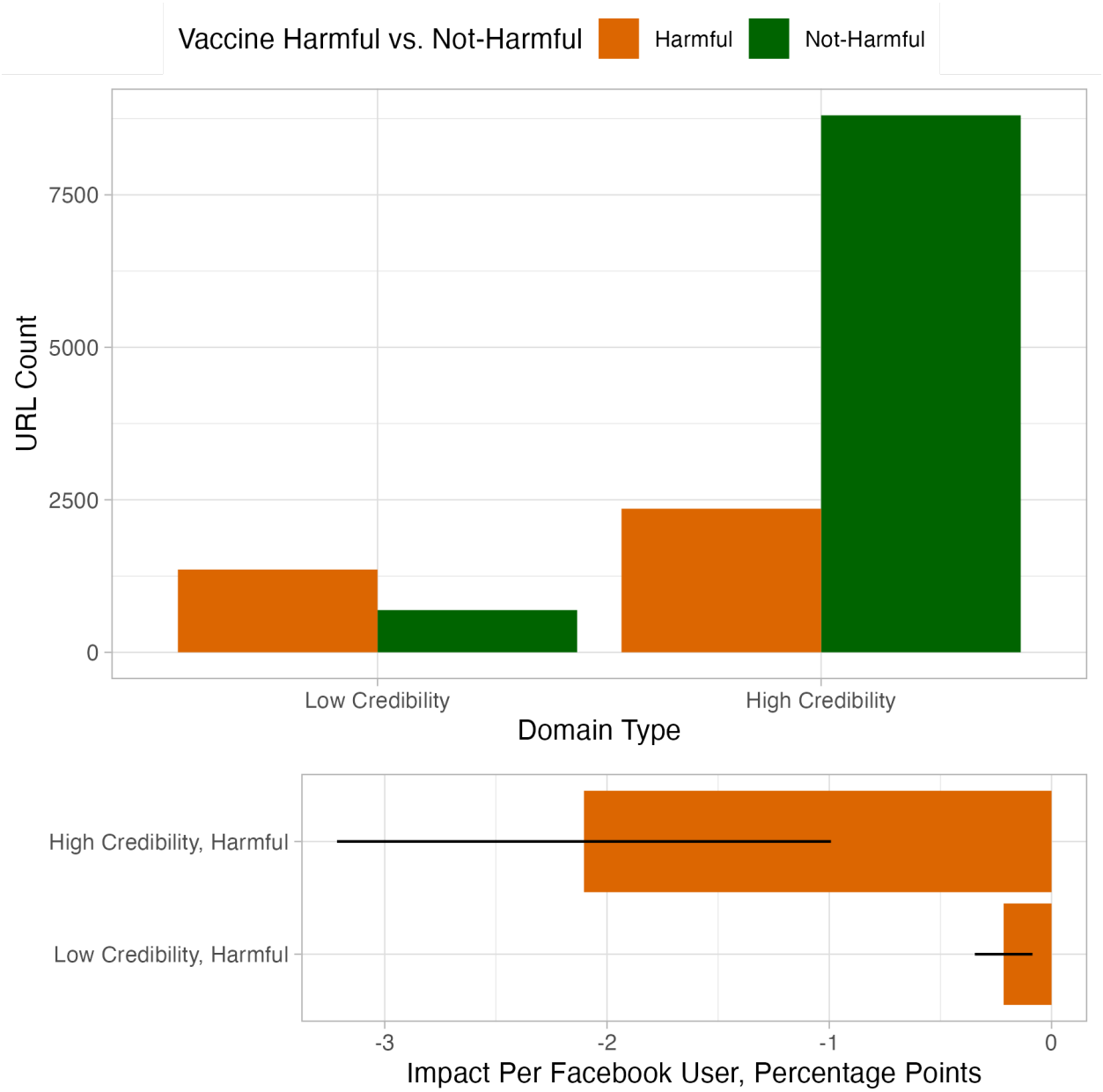

Figure S8: Relative harm by domain type. **Top** The counts of harmful vs. not-harmful vaccine URLs published by low vs. high credibility domains, respectively. A URL is labeled as harmful if it has a "Crowdsourced Aggregate Score" less than 3. **Bottom** The relative overall impact of harmful content from high vs. low credibility domains, where impact is calculated as the product of views and persuasive impact for each URL, summed over the set of all high and low quality URLs, respectively, and normalized by the number of Facebook users. 95% CIs are calcuated from a parametric bootrap of the model coefficients.

S4.6 Most Harmful Domains

In Figure S9, 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 harmful (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 harmful 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.

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."

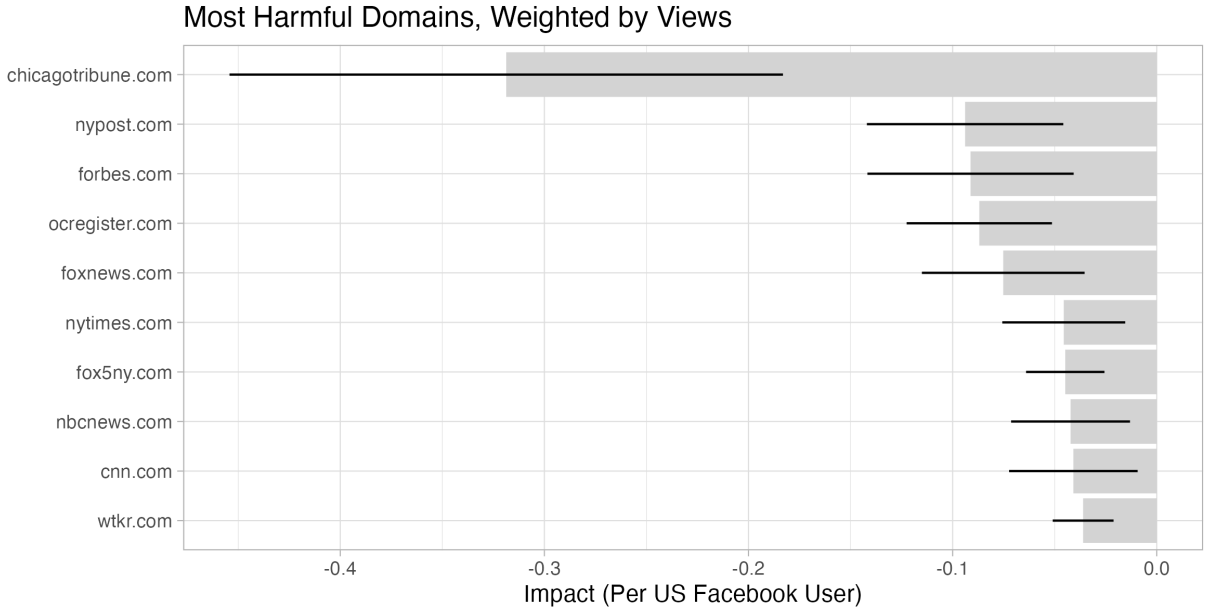

Figure S9: Top Harmful Domains, Weighted by Views

In Figure S10, we examine the top most harmful domains, ranked by the predicted persuasive impact of the average URL. Unlike Figure S9, this ranking does not weight impact by viewrs. Panel A shows the ranking over all domains, and Panel B shows the ranking for all domains with greater than 20 URLs.

Unlike Figure S10, 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.

S4.7 Subject-Level Heterogeneity

Here, 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 “Harmful" vaccine content compared to “Not Harmful" content,

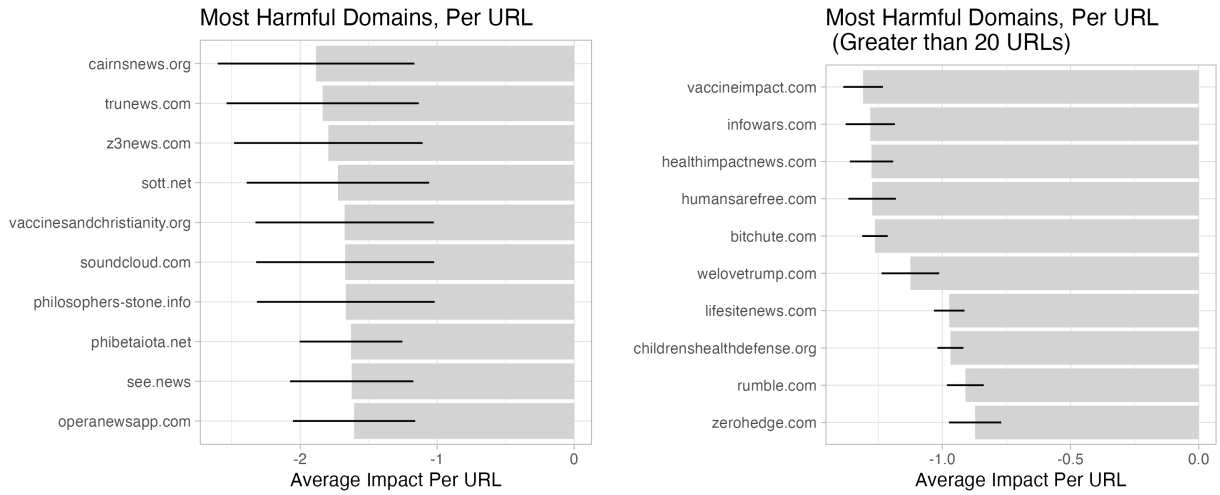

Figure S10: Top Harmful Domains, Average Impact Per URL

where we classify a URL as “Harmful" if it has a “Crowdsourced Aggregate Score" less than 3. Harmful content includes both vaccine-skeptical mainstream content as well as anti-vaccine misinformation. We report the proportion of vaccine content that is harmful, rather than the impact-per-user of harmful 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 harmful 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 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.

S4.8 Most Harmful Stories

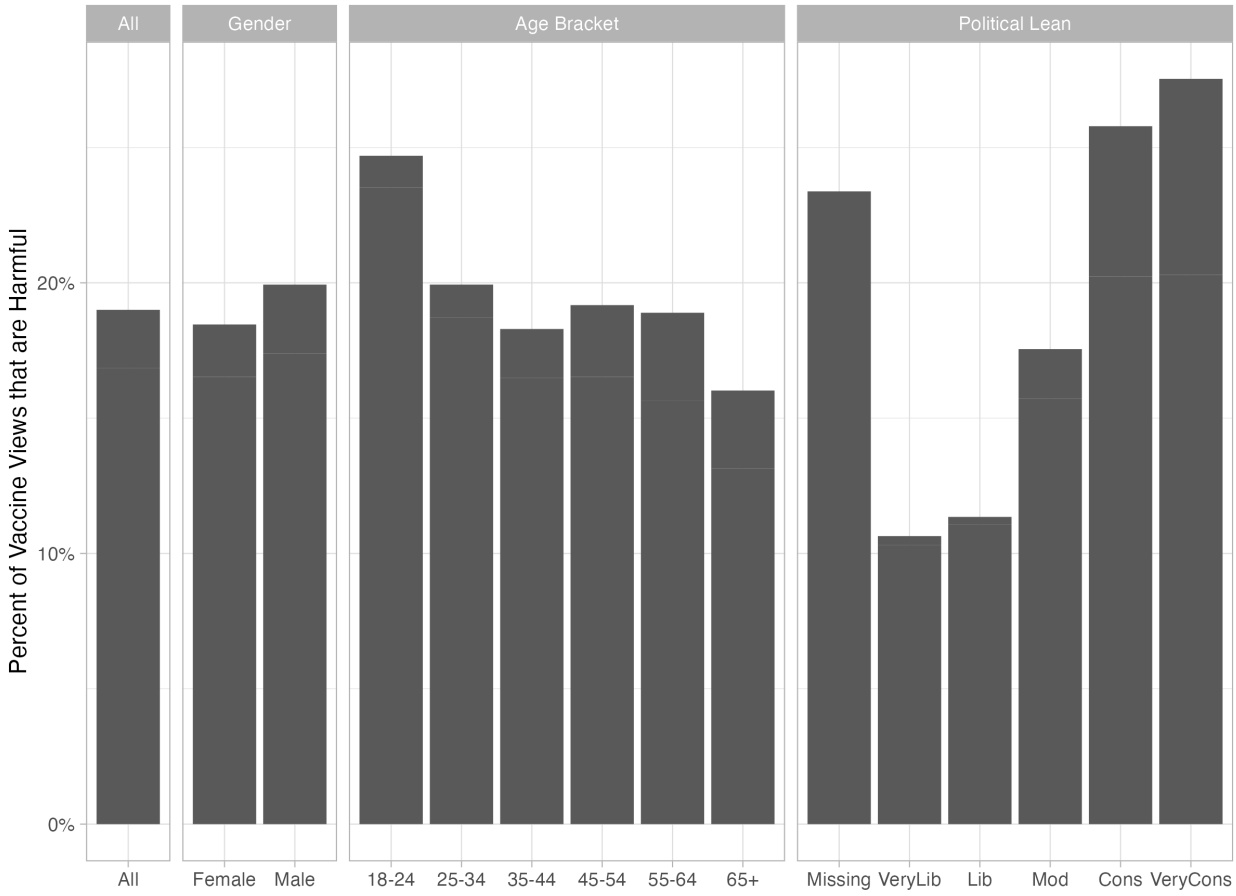

Figure S11: Percent of total vaccine views going to “Harmful" URLs for various demographics groups. URLs are classified as “Harmful" if they have a “Crowdsourced Aggregate Score" less than 3.

S4.9 Model, Lives Saved

We estimate the number of lives saved using the following calculation.

We define the following variables.

- P: The percentage point decrease in the vaccination rate due to vaccine-skeptical content on Facebook
- N_{FB}: The total number of US Facebook users in 2021
- V_l: The number of vaccinations needed to save one life

We assume the following values for these variables based on our results and external estimates:

$$\begin{aligned} P &= 2.3\,\% \text{ Based on our estimate in Figure 5C of the main text.} \\ N_{FB} &= 233,000,000 \text{ (users), [1]} \\ V_l &= 248 \text{ (vaccinations per life saved) [4]} \end{aligned}$$

The number of people who chose not to take the vaccine because of vaccine-skeptical content, $N_{affected}$, can be calculated as:

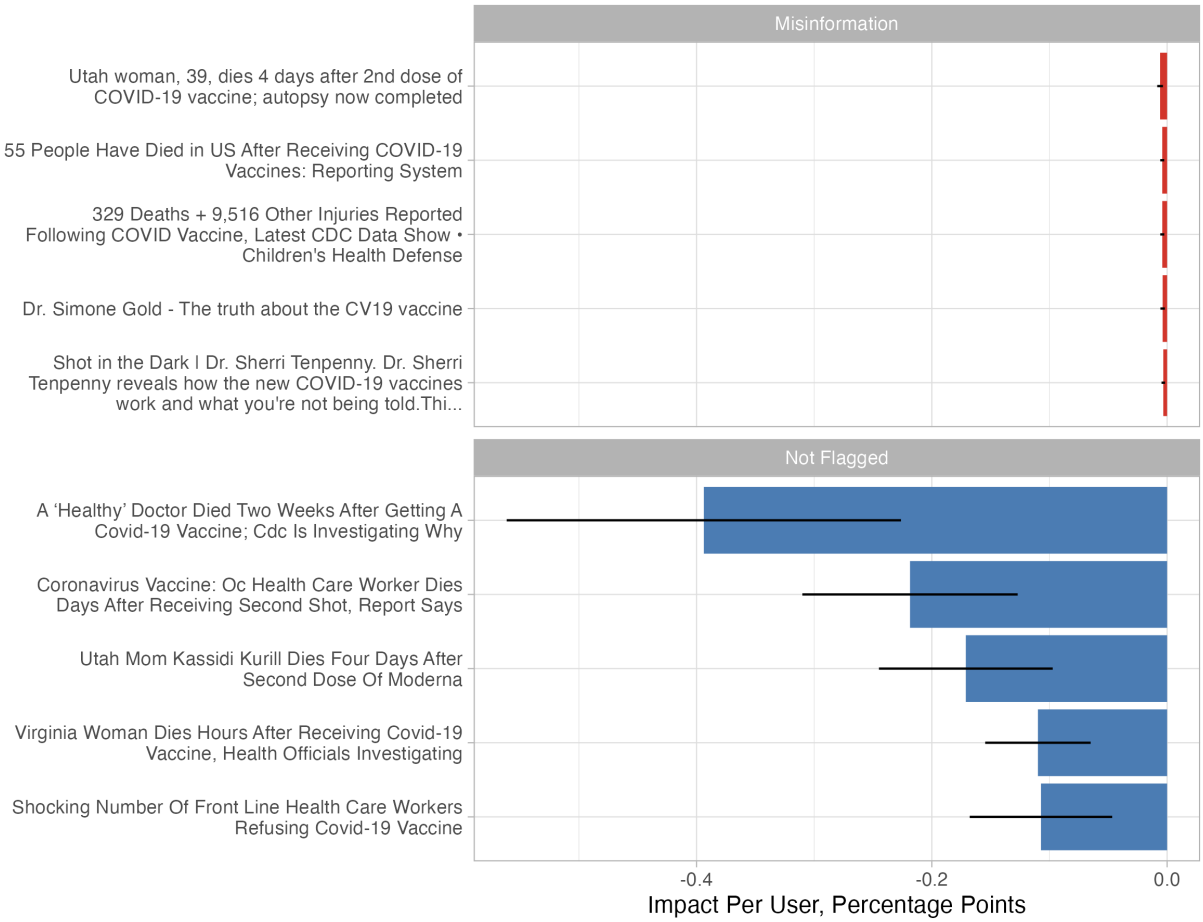

Figure S12: Most harmful stories, ranked by overall impact on vaccination intentions. Overall impact is calculated as the product of number of views times predicted persuasive effect, normalized by the number of US Facebook users. **Top** Top stories among misinformation flagged by fact-checkers as false, mixture of true and false, or missing context. **Bottom** Top stories among all content not flagged by fact-checkers.

$$N_{affected} = N_{FB} \times \frac{P}{100}$$

The number of lives that could have been saved, N_{saved}, can be calculated as:

$$N_{saved} = \frac{N_{affected}}{V_l}$$

Under these assumptions:

$$N_{affected} = 233,000,000 \times \frac{2.3}{100} = 5,359,000 \,\text{(people)}$$
$$N_{saved} = \frac{5,359,000}{248} \approx 21,606 \,\text{(lives)}$$

Therefore, we estimate that approximately 21,606 lives could have been saved if people had not been exposed to vaccine-skeptical content on Facebook.

S5 Robustness

S5.1 Contemporaneous Engagement Estimates

One drawback of our approach is that we combine mid-2022 experimental data and early-2021 exposure data, when ideally, we would have used contemporaneous estimates. For robustness, we show the results of the top most impactful stories among the headlines we experimentally tested in Study 2 using contemporaneous engagement data.

Our 90 stimuli in Study 2 were composed of highly-engaging mainstream and low-quality articles, collected using CrowdTangle. Although we do not have viewership data for these items, we do have the number of interactions (i.e. the number of shares, reactions, and comments) each article received on Facebook, which we use a proxy for viewership. We rank each article by its relative impact, calculated the product of the number of interactions it received and its estimated treatment effect. As can be seen in Figure S13, fact-checked true articles (most of which came from mainstream domains) had a much larger relative impact than false or misleading articles. These results suggest that our main findings, which show that little-scrutinized vaccine-skeptical content published by mainstream sources likely had a larger impact on lowering vaccination intentions than outright false content, are likely robust to the choice of timing.

S5.2 Contemporaneous Treatment Effect Estimates

As described in Section S4.4, we calculate our total impact estimates for URLs that we classify as “Harmful" based on whether they have a “Crowdsourced Aggregate Score" s_u less than 3. Content with $s_u >= 3$ are excluded from analysis.

In our original analysis, we find little evidence that content that promoted vaccination (i.e. with $s_u >= 3$) 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 [6, 11, 5]. 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 S12 estimates the net impact for promoting and harmful content at various cutoffs for whether or not a URL is considered “Harmful" vs. "Promoting", and for varying estimates of the effect for a single exposure to an item of vaccine-promoting content.

As an example, in Athey et al. [3] 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 S12. If we use a harmful 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.

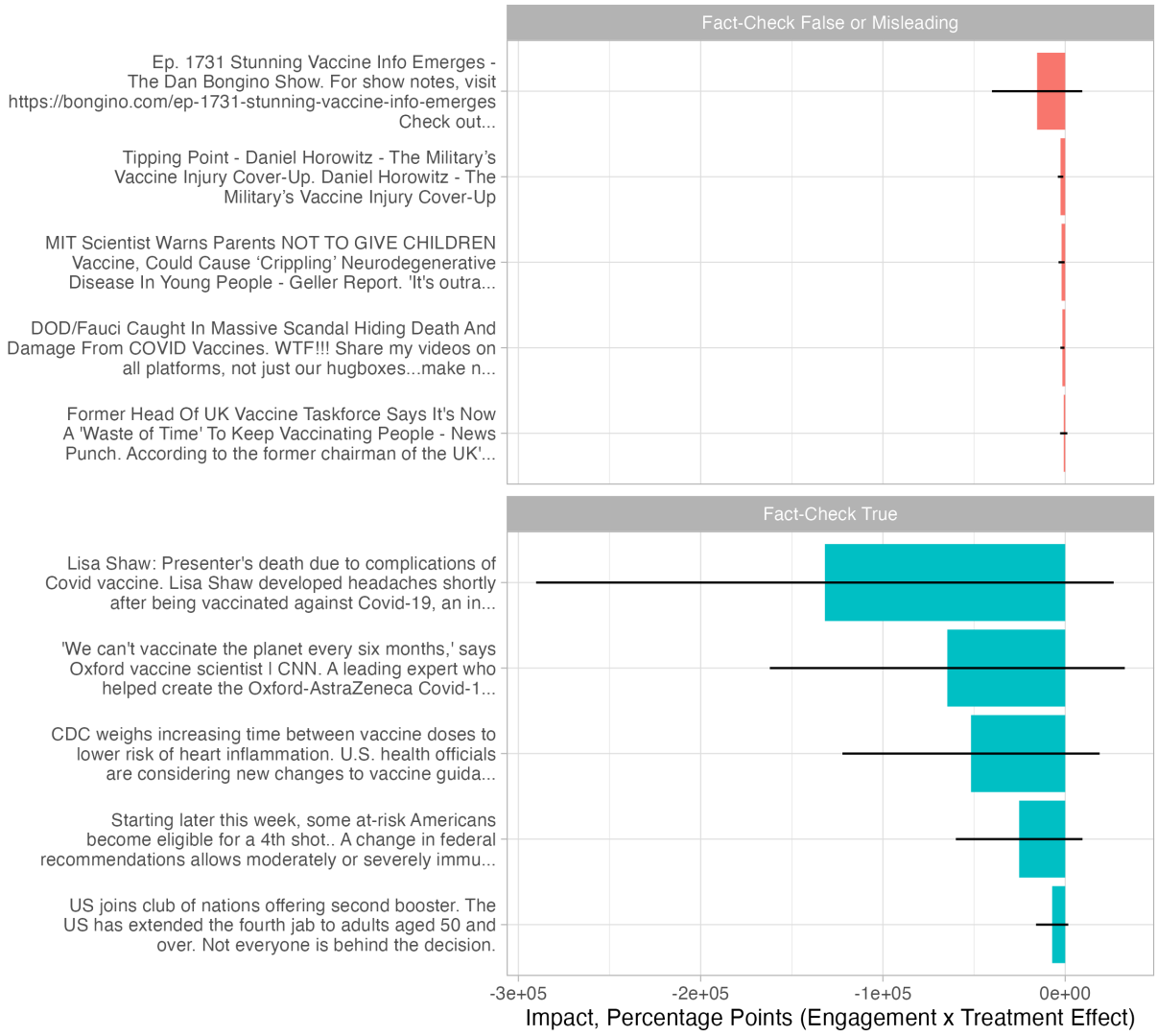

Figure S13: Most Influential Stories, Ranked by Interactions x Treatment Effect **Top** Articles rated by fact-checkers as false or misleading **Bottom** Articles rated by fact-checkers as true

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 [13]. This is 4.25X the size of the 1.5 percentage points average effect of misinformation on vaccination intentions we find in our experiments. Scaling the magnitude of our results by 4.25 would suggest that harmful misinformation and vaccine-skeptical 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 content on Facebook lowered vaccination intentions by 9.1pp based on more contemporaneous estimates of persuasive effect.

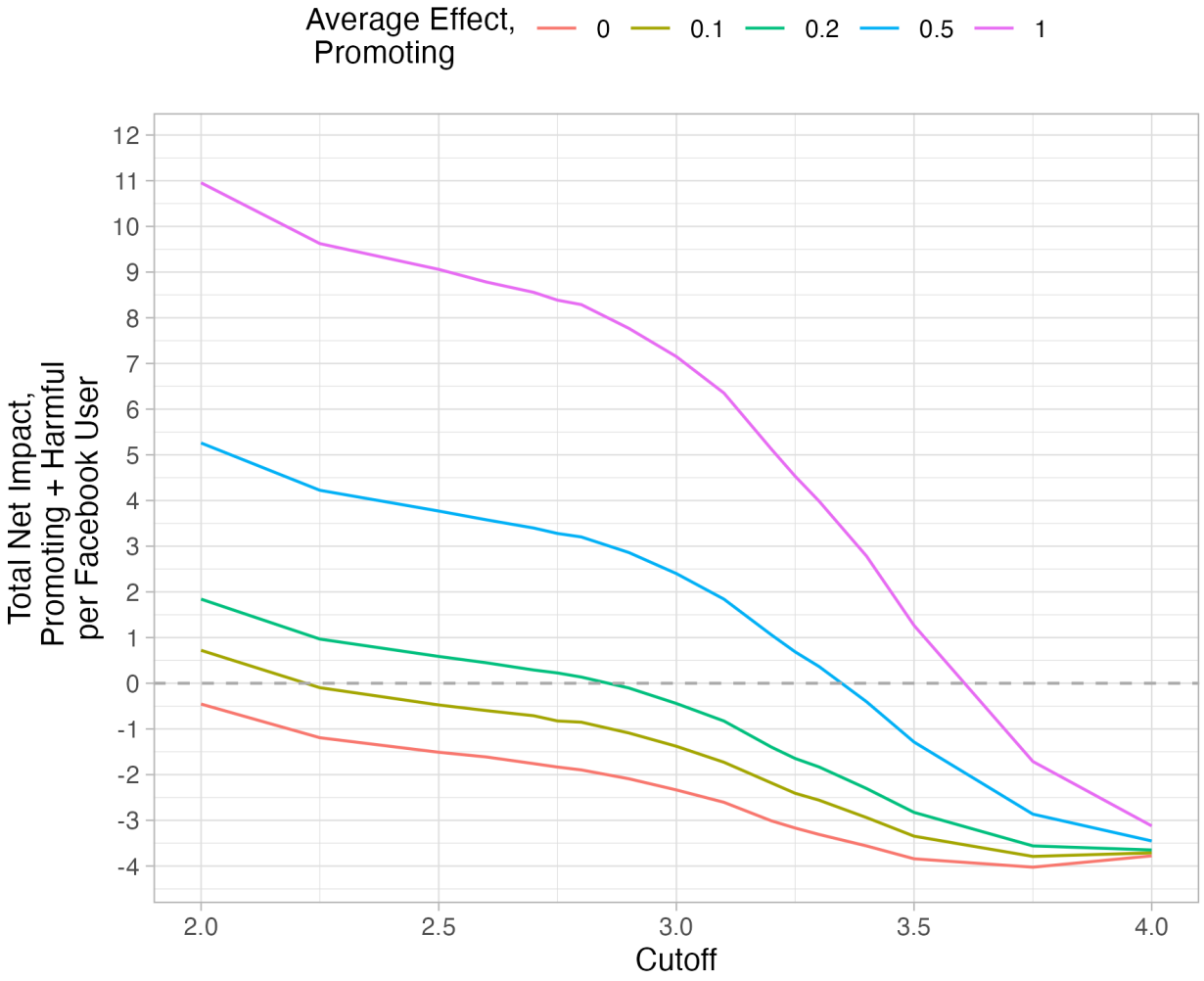

Figure S14: Net impact of promoting and harmful vaccine content as a function of the cutoff of the Crowdsourced Aggregate Score. “Cutoff" is the threshold of the score at which a URL is classified as either 1) harmful or 2) promoting. The colored lines show the net impact at different values for the average promoting effect.

References

[1] Number of Facebook users in the United States from 2018 to 2026 (In Millions), 2022. URL `https://www.statista.com/statistics/408971/number-of-us-facebook-users/`.

[2] J. Allen, A. A. Arechar, G. Pennycook, and D. G. Rand. Scaling up fact-checking using the wisdom of crowds. *Science advances*, 7(36):eabf4393, 2021. Publisher: American Association for the Advancement of Science.

[3] 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*, 120(5):e2208110120, 2023. Publisher: National Acad Sciences.

[4] R. J. Barro. Vaccination rates and COVID outcomes across US states. *Economics & Human Biology*, 47:101201, 2022. Publisher: Elsevier.

[5] D. Broockman and J. Kalla. When and Why Are Campaigns’ Persuasive Effects Small? Evidence from the 2020 US Presidential Election. 2020.

[6] J. G. Bullock. Partisan bias and the Bayesian ideal in the study of public opinion. *The Journal of Politics*, 71(3): 1109–1124, 2009. Publisher: Cambridge University Press New York, USA.

[7] K. P. Burnham and D. R. Anderson. Multimodel inference: understanding AIC and BIC in model selection. *Sociological methods & research*, 33(2):261–304, 2004. Publisher: Sage Publications Sage CA: Thousand Oaks, CA.

[8] A. Coppock. *Persuasion in parallel: How information changes minds about politics*. University of Chicago Press, 2023.

[9] A. Gerber and D. Green. Field Experiments: Design, Analysis, and Interpretation. New York: W. W, 2012.

[10] M. Harrer, P. Cuijpers, T. A. Furukawa, and D. D. Ebert. *Doing meta-analysis with R: A hands-on guide*. CRC press, 2021.

[11] E. Kamenica and M. Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011. Publisher: American Economic Association.

[12] J. Lasser, S. T. Aroyehun, A. Simchon, F. Carrella, D. Garcia, and S. Lewandowsky. Social media sharing of low-quality news sources by political elites. *PNAS Nexus*, 1(4):pgac186, Sept. 2022. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgac186. URL `https://doi.org/10.1093/pnasnexus/pgac186`.

[13] S. Loomba, A. de Figueiredo, S. J. Piatek, K. de Graaf, and H. J. Larson. Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa. *Nature human behaviour*, 5(3):337–348, 2021.