RESEARCH ARTICLE SUMMARY

MISINFORMATION

Quantifying the impact of misinformation and vaccine-skeptical content on Facebook

Jennifer Allen*, Duncan J. Watts, David G. Rand

INTRODUCTION: Researchers and public health officials have attributed much of the low uptake of the COVID-19 vaccine in the US to misinformation on social media. However, it is unclear whether misinformation had (i) the widespread exposure and (ii) the causal impact on vaccination intentions required to meaningfully alter the trajectory of US vaccination efforts. Moreover, content that raises questions about vaccines without containing outright falsehoods (which we term "vaccine-skeptical") might also play a role in driving vaccine refusal. In this work, we examine to what extent misinformation flagged by fact-checkers on Facebook (as well as content that was not flagged but is still vaccine-skeptical) contributed to US COVID-19 vaccine hesitancy.

RATIONALE: We posit that two conditions must be met for content to have widespread impact on people's behavior: People must see it, and, when seen, it must affect their behavior. That is, we define "impact" as the combination of exposure and persuasive influence.

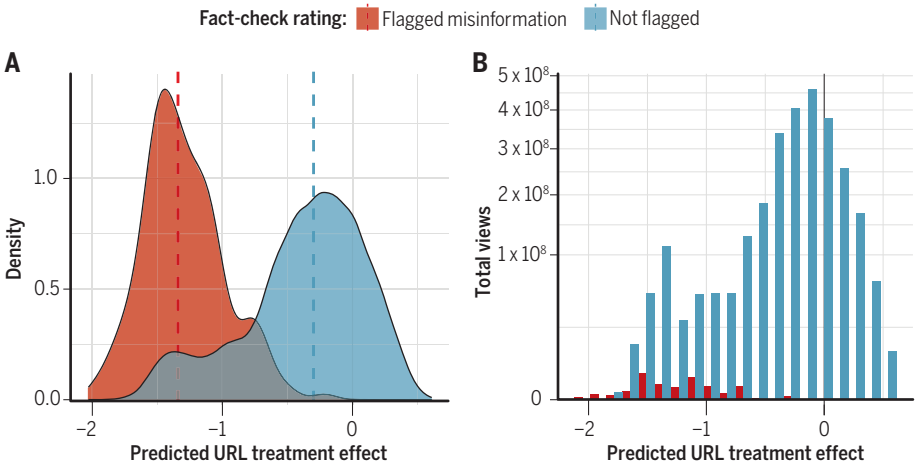
We apply this framework to quantify the impact that (mis)information on Facebook had on COVID-19 vaccination intentions in the US by combining experimental estimates of persuasive effects with Facebook exposure data. To begin, we conducted two experiments (total N = 18,725) on the survey platform Lucid measuring the causal effect of 130 vaccine-related headlines on vaccination intentions. We then used Facebook's Social Science One dataset to measure exposure to all 13,206 vaccine-related URLs that were popular on Facebook during the first 3 months of the vaccine rollout (January to March 2021). Finally, we developed a pipeline that incorporates the wisdom of crowds and natural language processing (NLP) to predict the persuasive effect of each Facebook URL from our survey estimates.

RESULTS: Analyzing our survey experiments, we found that while exposure to fact-checked misinformation can cause vaccine hesitancy, the degree to which a story implies health risks from vaccines best predicts negative persuasive influence. Our first experiment showed that misinformation containing false claims about the COVID-19 vaccine reduced vaccination intentions by 1.5 percentage points (P = 0.00004). Our second experiment tested both true and false claims and found that content suggesting that the vaccine was harmful to health reduced vaccination intentions, irrespectiv[illegible] any potential effect of the headline's verac[illegible].

Examining exposure on Facebook, we found that flagged misinformation URLs received 8.7 million views during the first 3 months of 2021, accounting for only 0.3% of the 2.7 billion vaccine-related URL views during this time period. In contrast, stories that were not flagged by fact-checkers but that nonetheless implied that vaccines were harmful to health—many of which were from credible mainstream news outlets—were viewed hundreds of millions of times.

We then used our crowdsourcing and NLP procedure to extrapolate the treatment effects of the 130 items to the 13,206 vaccine-related Facebook URLs. The URLs flagged as misinformation by fact-checkers were, when viewed, more likely to reduce vaccine intentions (as predicted by our model) than unflagged URLs. However, after weighting each URL's persuasive effect by its number of views, the impact of unflagged vaccine-skeptical content dwarfed that of flagged misinformation. Subsetting to those URLs predicted to induce hesitancy, we estimated that unflagged vaccine-skeptical content lowered vaccination rates by −2.28 percentage points {confidence interval (CI): [−3.4, −0.99]} per US Facebook user, compared with −0.05 percentage points (CI: [−0.07, −0.05]) for flagged misinformation—a 46-fold difference. Even though flagged misinformation had more of an impact when viewed, the differences in exposure were so large that they almost entirely determined the ultimate impact. For example, a single vaccine-skeptical article published by the *Chicago Tribune* titled "A healthy doctor died two weeks after getting a COVID vaccine; CDC is investigating why" was seen by >50 million people on Facebook (>20% of Facebook's US user base) and received more than six times the number of views than all flagged misinformation combined.

CONCLUSION: We find that flagged misinformation does causally lower vaccination intentions, conditional on exposure. However, given the comparatively low rates of exposure, this content had much less of a role in driving overall vaccine hesitancy compared with vaccine-skeptical content, much of it from mainstream outlets, that was not flagged by fact-checkers. Our work suggests that while limiting the spread of misinformation has important public health benefits, it is also critically important to consider gray-area content that is factually accurate but nonetheless misleading. ■



Impact of flagged misinformation versus unflagged content. (**A**) Distribution of 13,206 predicted URL treatment effects on vaccination intentions for flagged misinformation (red) versus unflagged content (blue). (**B**) The same histogram as in (A), weighted by the number of views each URL received on Facebook. Although misinformation has more negative persuasive effects, it is seen far less—and thus has a lesser impact—than unflagged content.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: jnallen@mit.edu
Cite this article as J. Allen *et al.*, *Science* **384**, eadk3451 (2024). DOI: 10.1126/science.adk3451

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.adk3451

RESEARCH ARTICLE

MISINFORMATION

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]

Low uptake of the COVID-19 vaccine in the US has been widely attributed to social media misinformation. To evaluate this claim, we introduce a framework combining lab experiments (total *N* = 18,725), crowdsourcing, and machine learning to estimate the causal effect of 13,206 vaccine-related URLs on the vaccination intentions of US Facebook users (*N* ≈ 233 million). We estimate that the impact of unflagged content that nonetheless encouraged vaccine skepticism was 46-fold greater than that of misinformation flagged by fact-checkers. Although misinformation reduced predicted vaccination intentions significantly more than unflagged vaccine content when viewed, Facebook users' exposure to flagged content was limited. In contrast, unflagged stories highlighting rare deaths after vaccination were among Facebook's most-viewed stories. Our work emphasizes the need to scrutinize factually accurate but potentially misleading content in addition to outright falsehoods.

In recent years, the spread of misinformation online has become a key concern for policy-makers and the public as well as 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 January 6th Capitol Hill riots and the rejection of public health messages during the COVID-19 pandemic. Despite a wealth of research on the spread of misinformation (*2*–*6*), its prevalence (*7*–*11*), and the psychology driving sharing of and belief in falsehoods (*12*–*15*), consideration of the real-world impact of exposure to misinformation [as opposed to impacts of larger algorithmic changes to platforms (*16*–*18*)] has been largely relegated to assertions in introductory paragraphs and discussion sections (*19*).

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—for example, President Joe Biden claimed that Facebook was "killing people" by allowing anti-vaccine misinformation to spread on the platform—little work has been done to show a causal connection (*20*). Numerous studies have shown a correlation between sharing and belief in social media misinformation and diminished COVID-19 vaccination (*21*–*28*). However, the causal direction of this correlational relationship is unclear. For example, other research has suggested that preexisting vaccine hesitancy inspires misinformation consumption rather than vice versa (*29*), whereas the few lab studies testing for a causal relationship between vaccine misinformation and behavioral intentions have shown conflicting evidence (*30*, *31*). 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," which we define as content that raises questions about vaccines but is not factually inaccurate, could play an important role in driving vaccine hesitancy (*29*, *32*, *33*).

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. And next, a large number of people must see it. Thus, impact arises from the interaction between exposure and persuasive influence: Harmful misinformation that no one sees does not make an impact, nor does misinformation that is widely seen but irrelevant to people's decision-making.

To estimate impact, we propose an approach that 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 the experiments using a combination of crowdsourcing and machine learning, we then estimate the overall impact of vaccine-related content shared on Facebook on vaccine hesitancy in the US. Critically, we model the impact of all popular vaccine-related URL content on Facebook, not only content that was flagged as misinformation by fact-checkers. 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.

Although the term "misinformation" has been defined in many ways by different scholars (*1*, *34*–*36*), researchers in fields such as computer science, psychology, and political science (as well as technology company moderation policies) often focus on URL content (i.e., "news") that has been flagged as false or misleading by professional fact-checkers (*19*, *37*–*41*). Here, we adopt the same convention. We then designate all other URL content that does not meet this definition, but that nonetheless might induce vaccine hesitancy, as "vaccine-skeptical." This distinction sets up our key research question: How does the impact of flagged misinformation content, which receives substantial attention from researchers and platforms, compare with the impact of vaccine-skeptical content, which has been far less scrutinized?

To answer this question, we apply our framework, which estimates impact as the interaction of persuasive influence and exposure. The paper is organized into four sections. First, we analyze the results of two survey experiments and show that, although exposure to fact-checked misinformation can cause vaccine hesitancy, the degree to which a story implies health risks from vaccines, rather than veracity, best predicts negative persuasive influence. Next, we analyze exposure to popular Facebook vaccine content from early 2021 and find that flagged misinformation gained relatively little traction on Facebook compared with unflagged stories (largely from credible mainstream news outlets) that nonetheless implied that vaccines were harmful to health. Then, we develop a methodology to predict the persuasive effect of this popular Facebook content by leveraging a combination of the "wisdom of crowds" and machine learning to generalize the causal effects measured in our experiments. Finally, we combine the exposure data from Facebook with the results of our predictive model and find that the overall predicted impact of vaccine-skeptical content on vaccine hesitancy dwarfs that of anti-vaccine content flagged as misinformation by fact-checkers. Together, these results suggest that policies that prioritize vetting content with obvious factual inaccuracies target only a small minority of content that could lead to public health harms.

[1]Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA. [3]Annenberg School for Communication, University of Pennsylvania, Philadelphia, PA, USA. [4]Operations, Information, and Decisions Department, University of Pennsylvania, Philadelphia, PA, USA. [5]Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA, USA. [6]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA.
*Corresponding author. Email: jnallen@mit.edu

Method and results

Persuasive effects of COVID-19 vaccine (mis)information

First, we consider which types of vaccine content changed willingness to take a COVID-19 vaccine, conditional on exposure (i.e., after being viewed), using two large-scale online survey experiments. Our approach differs from other misinformation studies in that we (i) measured behavioral intentions as our outcome of interest, rather than belief in or sharing of false (versus true) claims; and (ii) assessed differences in persuasive effects across messages, irrespective of their veracity. We emphasize that our approach distinguishes veracity and persuasive impact, which are not inherently related but can be conflated in misinformation research. For example, "COVID-19 is only as deadly as the seasonal flu" and "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 only the former is likely to affect vaccine intentions (*42*). Conversely, factually accurate content might also affect vaccine intentions, for example, news of the government pausing rollout of the Johnson & Johnson COVID-19 vaccine because of concerns about blood clot risks (*43*).

In study 1, N = 8603 American participants from the online survey platform Lucid were shown a neutral control post or a single piece of vaccine misinformation from a set of 40 articles, videos, and posts previously debunked by fact-checkers [see supplementary materials (SM) section S0.2.3 for details on stimuli]. To assess impact, participants were asked to answer a set of questions regarding their willingness to take a COVID-19 vaccine (which we combined into a COVID-19 vaccination index) before and after exposure. Consistent with conventional wisdom, we found that exposure to a single piece of vaccine misinformation decreased vaccination intentions by 1.5 percentage points on average (P = 0.00004). This effect did not vary significantly on the basis of participants' pretreatment vaccination intentions, gender, age, political party, or vaccine status (P > 0.2 for all after Benjamini-Hochberg correction; see SM sections S1.5.4 and S1.5.5 for details). Nonetheless, it did vary substantially across different pieces of misinformation: Whereas misinformation items in the bottom decile of the distribution had double the average effect—lowering vaccination intentions by 3 percentage points—the stimuli in the top decile had a treatment effect of zero (see SM section S0.2.7). In other words, an item did not lower vaccination intentions simply by virtue of being false, which suggests that other dimensions of the content were relevant beyond veracity.

In study 2, we moved beyond study 1's focus on fact-checked misinformation by collecting a representative set of 90 highly shared vaccine-related articles sampled from Facebook, balanced across topic and domain quality (see SM section S0.2.3 for details on stimuli). 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 in 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 (i) surprising, (ii) plausible, (iii) favorable to Democrats versus Republicans, (iv) familiar, and (v) whether the item suggested that the vaccine was harmful versus helpful to a person's health (see SM section S0.2.4 for details on ratings). 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 SM section S1.5.2 for details, including disaggregated analyses).

We found that the only content dimension that consistently predicted a headline's effect on vaccination intentions was the extent to which the headline suggested that the vaccine was harmful to a person's health (Fig. 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 −0.69 percentage points (SE: 0.19, P = 0.0003) for a model with just harmful-to-health as a predictor and −0.49 percentage points (SE: 0.23, P = 0.036) for a model including other potential predictors (see SM section S1.5.3 for associations with other content dimensions). We found no significant effect on vaccination intentions of whether the headline came from a low-quality domain (e.g., childrenshealthdefense.org, a site known for spreading anti-vaccine disinformation) as opposed to a mainstream domain (e.g., nytimes.com) (β = −0.27, SE: 0.23, P = 0.24). Falsity (as judged post hoc by professional fact-checkers; see SM section S0.2.4 for details) was associated with a more negative effect on vaccination intentions (β = −0.85, SE: 0.27, P = 0.002). Perhaps unsurprisingly, false claims were more likely to suggest that the vaccine is harmful to health (β = 0.77, SE: 0.08, P < 0.00001, as can be seen in Fig. 1). However, when predicting treatment effect size using both veracity and the extent to which the headline suggested that the vaccine was harmful to health, as well as their interaction (with variables z-scored for interpretability), harmful-to-health remained significant (β= −0.38, P = 0.005), whereas veracity did not (β= −0.21, P = 0.17; there was also no significant interaction, β= −0.19, P = 0.16). These results indicate that suggesting the vaccine was harmful to health reduced vaccination intentions, irrespective of any potential effect of whether the headline was factually inaccurate.

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

We next examined levels of exposure to vaccine-related content on Facebook. Some prior research supports the "infodemic" framing, identifying cases where viral COVID-19 vaccine misinformation shared by a small number of misinformation "superspreaders" generated millions of interactions on social media (*44*–*47*). 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 little 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 (*48*). Specifically, we identified 13,206 URLs about the COVID-19 vaccine shared publicly >100 times on Facebook and published during the first 3 months of 2021 (the initial rollout period for the vaccine in the US; see SM section S0.1.1 for dataset details).

We found that URLs flagged by professional fact-checkers as false, out-of-context, or a mixture—which we will refer to as "flagged misinformation" in our subsequent analyses—received 8.7 million views, accounting for only 0.3% of the 2.7 billion vaccine-related URL views during this time period (Fig. 2A). Similarly, content from domains rated as low-credibility [as determined in (*49*)] received only 5.1% of views. Thus, exposure to flagged URL misinformation about vaccines on Facebook was relatively infrequent, owing to some combination of low baseline user viewership and explicit downranking by Facebook (*7*, *10*, *50*).

As just noted, however, even content not flagged by fact-checkers may have negative effects on vaccination intentions [we will refer to content that is not flagged as misinformation but still raises questions about the vaccine's safety and effectiveness as "vaccine-skeptical" (*29*)]. For example, examining the top 10 most-viewed vaccine-related story clusters in the dataset revealed that several articles published by mainstream news organizations did cast doubt on the vaccine. For example, the most-viewed URL across all 13,206 URLs during

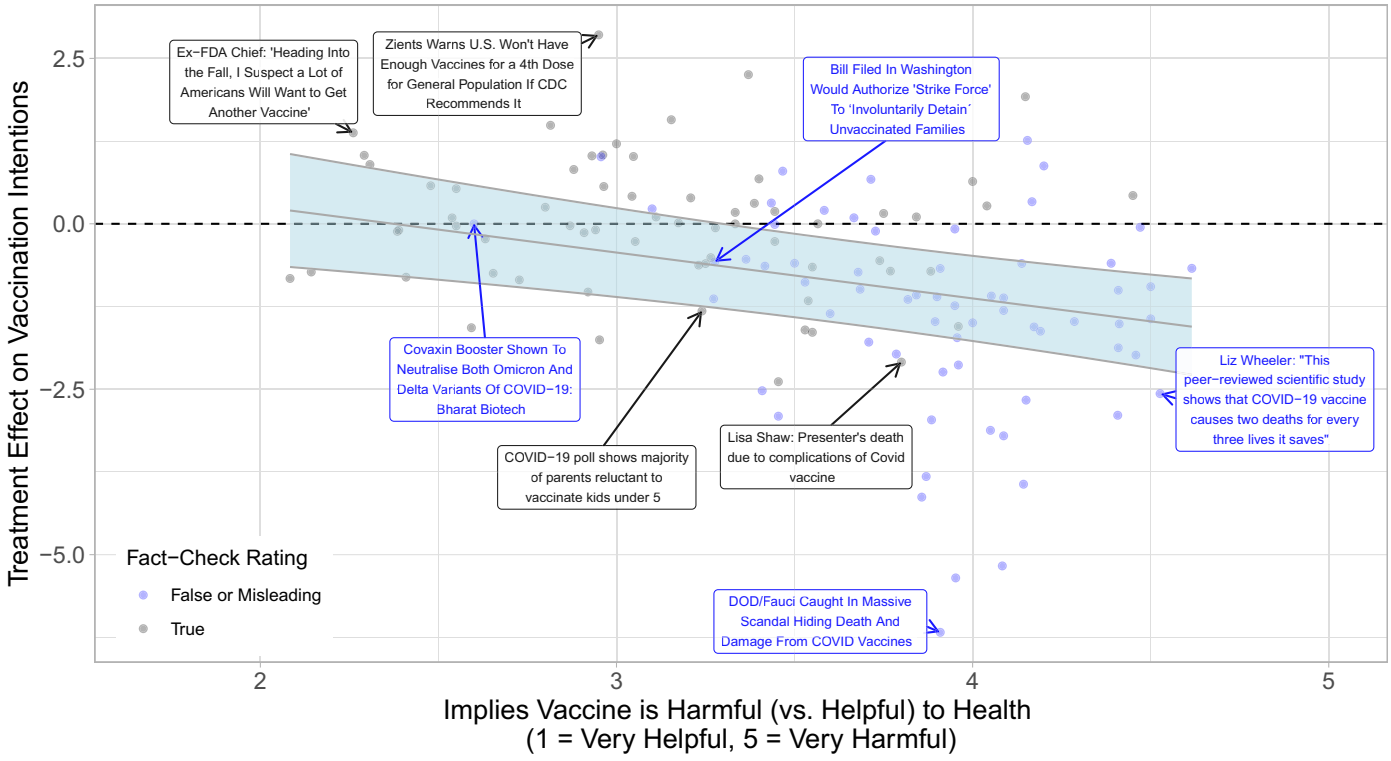


Fig. 1. Effect of vaccine-related headlines on vaccination intentions as a function of perceived harm. False or misleading articles are indicated in blue, and factually accurate articles are indicated in black. Overlaid in gray is the best-fit line and 95% CI from a random-effects meta-regression with 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.

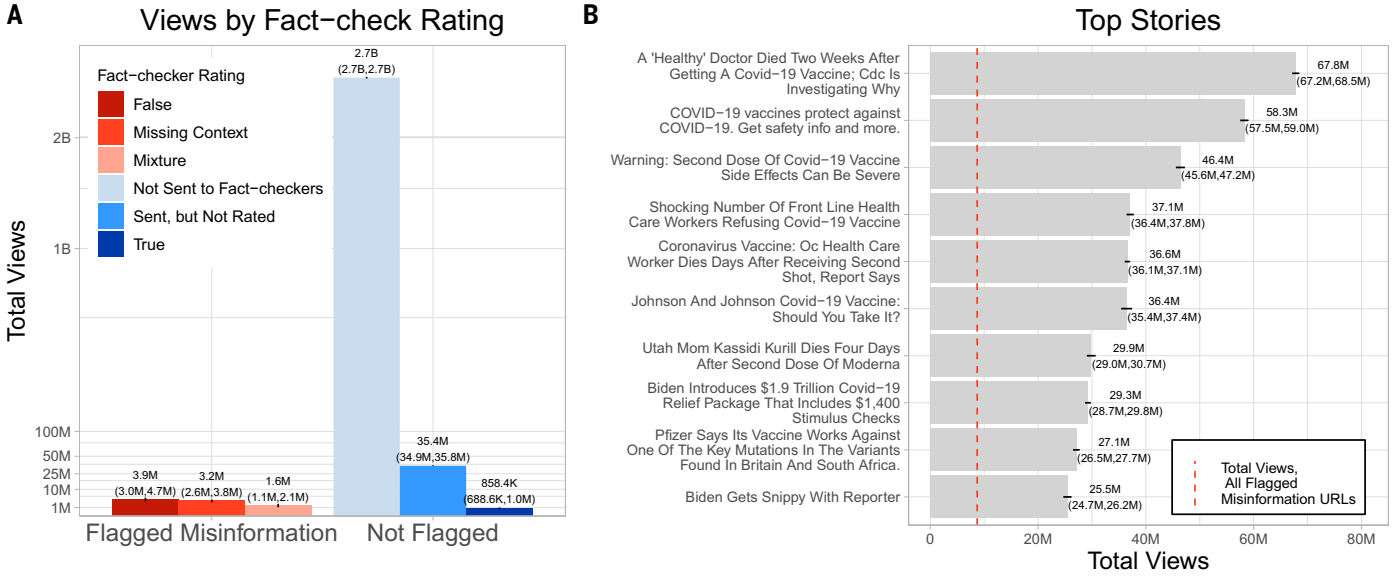


Fig. 2. Exposure to vaccine-related content on Facebook that was publicly shared >100 times on Facebook during the first 3 months of 2021. View counts are shown with 95% CIs to account for differentially private noise (see SM section S0.1.1 for more information). (**A**) Total views for misinformation versus 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 vaccine-related views during this time period. (**B**) 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 on the basis of their headlines and descriptions (see SM section S0.1.2 for clustering details). The aggregate number of views on all misinformation URLs is indicated by the red dashed line.

this time period was a *Chicago Tribune* article titled "A healthy doctor died two weeks after getting a COVID vaccine; CDC is investigating why." This URL was seen by 54.9 million people on Facebook (>20% of Facebook's US user base), and all URLs related to this story were seen at least 67.8 million times (more than six times the number of views on all flagged misinformation combined).

This news story and others like it contained no intentional falsehoods and, in many cases, indicated the uncertainty surrounding the true cause of death (at least in the body of the article). Nonetheless, the story's clear implication (especially from the clickbait-style headline) was that the vaccine may have been harmful to health. Although some scholars and journalists have identified these kinds of headlines as true but misleading (e.g., "missing context") and argue that they should be considered misinformation (*33*, *51*), they are substantively different from—and much more ambiguous than—outright "fake news" and otherwise fact-checked false stories that academic and journalistic attention has focused on since the 2016 election. Furthermore, little work has been done to identify these ambiguous stories in a systematic way at scale, compared with tracking outright falsehoods or content from low-credibility sources. This lack of scrutiny is meaningful: We find that exposure to vaccine-skeptical content far outstripped exposure to flagged misinformation.

Scaling up estimates of persuasion conditional on exposure

In this section, we introduce an approach for identifying potentially harmful content at scale—by predicting its persuasive effect rather than on the basis of its veracity or source credibility. To estimate the impact of flagged misinformation and vaccine-skeptical content on Facebook, we needed 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 used a combination of crowdsourcing and machine learning to generalize the results from our survey experiments to the full Facebook URL dataset.

To begin, we recruited crowd raters from CloudResearch'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 (see SM section S0.3.1 for details). We then created a "crowdsourced aggregate score" by averaging this measure with the previously discussed crowdsourced ratings of (i) whether the headline suggests that vaccines are harmful or helpful to health and (ii) headline accuracy [which, consistent with prior work (*52*–*54*), agreed highly with expert judgments; correlation coefficient (r) = 0.72, $P < 0.000001$]. Using a random-effects meta-regression to predict each of the 130 headlines' observed causal effect on vaccination intentions in studies 1 and 2, we found that this crowdsourced aggregate score is predictive of the actual treatment effects (Fig. 3; correlation, adjusted for sampling error: 0.75, $P = 0.0001$, pseudo-$R^2 = 0.34$; see SM sections S0.3.2 and S2 for details). These results demonstrate that while the crowd might not predict a given individual treatment effect with high accuracy (owing 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. As 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 treatment effects.

Next, we recruited additional raters and had them rate 1139 of the 13,206 URLs from Facebook (oversampling URLs that were highly viewed, covering diverse events, and flagged by fact-checkers; see SM section S0.3.4). We randomly split our labeled data into an 85/15 train-test split on our labeled data, stratified on the crowdsourced predicted effect. We then trained a machine learning model using a COVID-Twitter-BERT architecture (*55*) to predict the crowdsourced scores for the full set of 13,206 URLs. We found that our model is capable of predicting the crowdsourced aggregate score in our holdout test set; the vast majority—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 simpler binary classification task predicting whether the URL was hesitancy-inducing (which we operationalize as being below the aggregate score midpoint; see SM section S3.6 for discussion of cutoff), the model had a 97% area under the receiver operating characteristic curve (AUC), 91% accuracy, and a 4% false-positive rate [i.e., only 4% of URLs the crowd thought would increase support for vaccine hesitancy were predicted by our model to decrease support; see SM section S3.1 for details and robustness checks]. To predict the

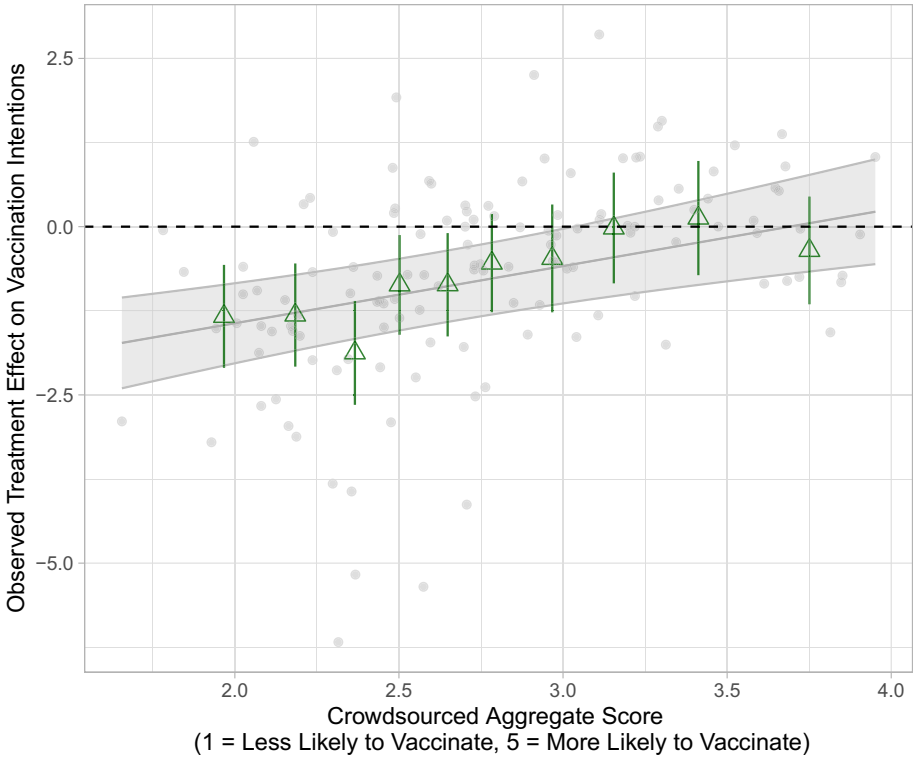


Fig. 3. Treatment effect on vaccination intentions as a function of the crowdsourced aggregate score. The crowdsourced aggregate score was calculated using the crowds' (i) prediction of whether the story would increase or decrease willingness to vaccinate, (ii) harmful- versus helpful-to-health rating, and (iii) accuracy rating (see SM section S2.1 for details). Each point corresponds to one of the 130 items in studies 1 and 2. The overlaid gray line is the best-fit line and 95% CI 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 consider all items below the midpoint of 3 to be "hesitancy-inducing."

causal effect (conditional on exposure) of each URL, our ultimate goal, we passed these predicted crowdsourced aggregate scores into the meta-regression model in Fig. 3 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 found that the estimated effect on vaccination intentions of the average flagged misinformation URL is substantially more negative than the average URL not flagged as misinformation [$t_{197} = -40.2$, $P < 0.00001$); Fig. 4A]. The median flagged URL had an estimated treatment effect of −1.36 percentage points {95% confidence interval (CI): [−1.91, −0.73]}, more than four times larger in magnitude than the estimated treatment effect of the median unflagged URL (−0.3 percentage points, 95% CI: [−0.92, 0.34]). That is, when seen, the typical fact-checked misinformation URL was predicted to reduce vaccination intentions much more than unflagged content.

Quantifying harm caused by hesitancy-inducing (mis)information on Facebook

Does the significantly larger negative effect of flagged misinformation relative to unflagged URLs, conditional on exposure, imply that flagged misinformation had an outsized impact on Facebook during the vaccine rollout? To answer this question, we must combine the estimated treatment effects shown in Fig. 4A with the exposure data from the Social Science One dataset shown in Fig. 2.

When we weight the estimated treatment effect for each URL by the number of views that URL received (Fig. 4B), we see a very different picture: The impact of the URLs flagged as misinformation is dwarfed by the impact of URLs that were not flagged but that were predicted to increase hesitancy. These unflagged URLs had a somewhat weaker predicted impact when viewed than flagged posts, but they were viewed by vastly more people.

Because our research question focuses on vaccine refusal, we then subsetted to the 3711 URLs predicted to be hesitancy-inducing (see SM section S4.4 for an analysis of the full dataset of 13,206 URLs, which gives qualitatively similar results, and SM section S5.2 for discussion of potential impact of pro-vaccine content). We differentiate between URLs flagged as misinformation versus unflagged but hesitancy-inducing URLs. As noted previously, we use the term "vaccine-skeptical" to refer to these unflagged hesitancy-inducing URLs. Also as described above, flagged misinformation URLs received only a small percentage of the total views of hesitancy-inducing content—98% of the >500 million views of content estimated to reduce vaccinations was vaccine-skeptical rather than flagged misinformation. Therefore, when we take the product of exposure and estimated impact for each hesitancy-inducing URL (and normalize by the total number of US Facebook users for interpretability), we estimate that this vaccine-skeptical content not flagged by fact-checkers lowered self-reported vaccination intention by a predicted −2.28 percentage points (CI: [−3.4, −0.99]) per US Facebook user, compared with an effect of only −0.05 percentage points (CI: [−0.07, −0.02]) for flagged misinformation (Fig. 4C). Taking the ratio of the two point estimates, we find that the estimated impact of vaccine-skeptical content was 46-fold greater than that of flagged misinformation.

To shed more light on the vaccine-skeptical content in our data, we examined the stories predicted to have the largest effect on vaccination intentions—none of which were flagged by fact-checkers (see fig. S14). We found that coverage of young, healthy people's deaths after vaccination—with headlines that did not contextualize how exceedingly rare such deaths were, or the uncertainty of the vaccine's role in causing these deaths—achieved disproportionate reach, and therefore had a disproportionate estimated impact, during this time period.

Critically, many of these high-impact vaccine-skeptical articles came from mainstream sources. Although a much larger fraction of content from low-credibility domains (66%) was

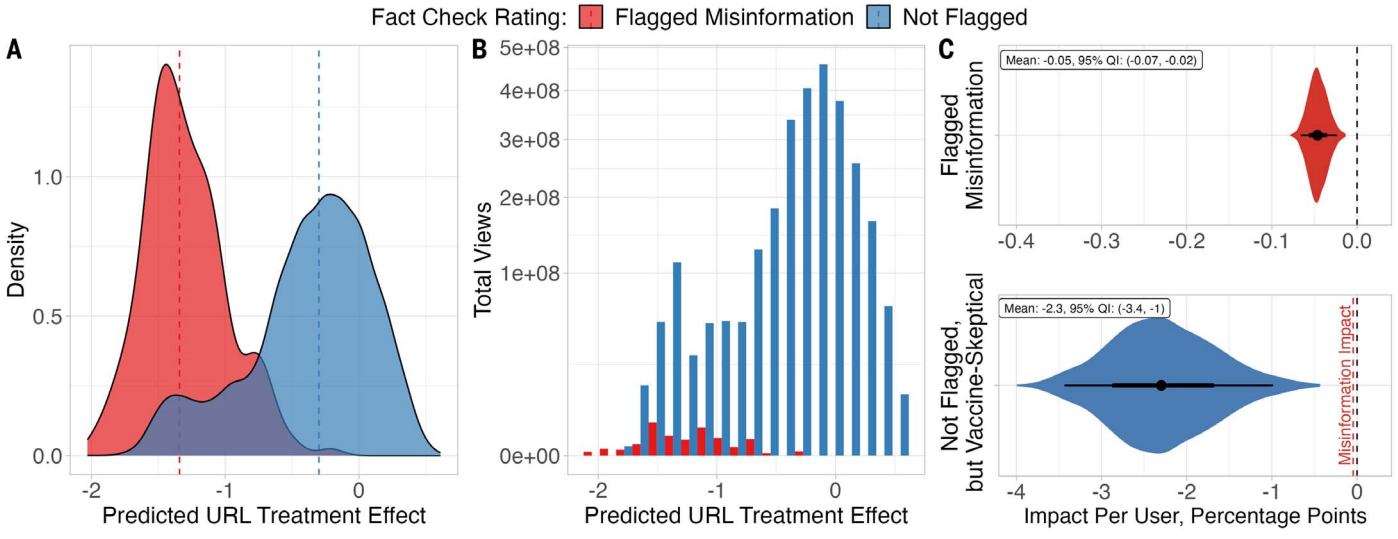


Fig. 4. Predicted impact on vaccination intentions of vaccine-related URLs that were publicly shared >100 times on Facebook during the first 3 months of 2021. (**A** and **B**) Distribution of predicted treatment effect on vaccination intentions across all URLs, comparing 186 URLs flagged as misinformation (shown in red) versus 13,020 URLs not flagged by fact-checkers (shown in blue). (A) Density plots for predicted treatment effects. Dashed lines represent the medians of the distributions. (B) The same histogram of URL treatment effects as in (A), weighted by the number of views each URL received. Note that the *y* axis in (B) is shown on a square-root scale for better visualization. (**C**) Overall predicted treatment effect among the 3711 hesitancy-inducing URLs (i.e., predicted crowdsourced aggregate score below scale midpoint), comparing the 183 URLs flagged as misinformation versus the 3528 URLs that were not flagged (which we refer to as vaccine-skeptical). Shown is the total impact across each type of URL, normalized by the number of US Facebook users. The point estimates (in black) are shown with 50 and 95% CIs, calculated from a parametric bootstrap of our coefficients. We additionally compute analytical prediction intervals, assuming worst-case correlation among errors, and find that our results are robust even under these extreme assumptions (see SM section S5.5). Note that for readability, the scales for flagged misinformation differ from vaccine-skeptical content; in the vaccine-skeptical panel, we label the average impact for flagged misinformation with a red dashed line for reference.

estimated to meaningfully reduce vaccination intentions compared with 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 estimated decrease in vaccination intentions (see SM sections S4.5 and S4.6 for more details). Despite the worries about misinformation superspreaders (*56*), mainstream news outlets, such as the *New York Post* and Fox News, as well as local news outlets, were the sources of URLs that had the biggest overall negative predicted impact on vaccination intentions (see fig. S12 for full list of most hesitancy-inducing domains).

Discussion

Here, we have introduced a method combining crowdsourcing, machine learning, and large-scale observational data for estimating the causal effect of social media on societal outcomes. Using this method, we have investigated the effect of COVID-19 vaccine-related URLs that circulated on Facebook in early 2021 on self-reported vaccine hesitancy, estimating that the combination of flagged misinformation and unflagged vaccine-skeptical content lowered US vaccination intentions by 2.3 percentage points per Facebook user. However, contrary to conventional wisdom, we show that this effect was driven almost entirely by vaccine-skeptical content from mainstream sites that was not flagged as misinformation by fact-checkers, rather than by outright false content published by fringe outlets.

What do these findings imply about the efficacy of the most common interventions for identifying and fighting misinformation? The typical approaches identify misinformation using third-party fact-checker labels or ratings of domain quality. They involve strategies such as surfacing fact-checker labels or corrections, penalizing low-quality domains, or scaling digital literacy interventions that advise "checking the source" of content (*2*, *57*–*62*). Even automated systems designed to detect and limit the spread of fake news online primarily use databases of fact-checked claims as training data (*63*–*66*). Although these veracity-oriented interventions may have reduced exposure to content that was harmful when viewed, they are unlikely to have reduced the spread of the type of content identified as having the most overall negative impact in our analyses: unflagged vaccine-skeptical stories often published by mainstream outlets, including the Pulitzer Prize–winning *Chicago Tribune*. Had exposure to this content been prevented, we estimate that vaccination intentions could have been 2.3 percentage points higher on average among Facebook's 233 million US users—translating into ~3 million more vaccinated Americans [assuming that effects on actual vaccinations rates were 60% the magnitude of effects on vaccination intentions (*67*)]. It has been estimated that 248 additional vaccinations translate into one additional life saved (*68*), implying that many lives could have been saved had vaccine-skeptical content not been published or allowed to spread unchecked on Facebook.

Our work also has important limitations. First of all, our survey experiments and our observational data come from different time periods. The Facebook viewership data (which are only available several months after the views occurred) are 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 were applied by technology companies). To help address concerns regarding the delay in the present data, we performed several robustness checks which reexamined our results with contemporaneous data for experimental effects and engagement (as a proxy for exposure), respectively, which show similar patterns (see SM sections S5.1 and S5.2).

Our work also measures survey intentions to take a COVID-19 vaccine, rather than actual vaccination behavior, and thus could be overstated. Reassuringly, Athey *et al.* (*67*) found that survey and behavioral measures of COVID-19 vaccination are substantially correlated and, in particular, found that a 1 percentage point increase in vaccination intentions measured through Facebook surveys corresponded to a 0.6 percentage point increase in the actual county-level vaccine uptake rate (*67*). Other work has also found a substantial correlation between individual intentions to vaccinate and vaccine uptake for non–COVID-19 vaccines (*69*–*71*), and a meta-analysis shows a 0.55 ratio between effects on intentions and behavior across a variety of interventions (*72*). Furthermore, the intention–behavior gap is largely concentrated among those who intend but fail to take an action, and there is less theoretical evidence to suggest that intentions to avoid an action, as in the case of vaccine refusal, would be subject to an equivalently sized gap (*73*, *74*). Understanding more quantitatively how the causal effects that change vaccination intentions we estimate here translate into actual vaccine uptake—for example, by linking variation in exposure to vaccine-related media to regional vaccination rates—is an important area for future work.

Furthermore, our Facebook data included only URL 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 possible that misinformation (compared with factual information) was relatively more prevalent among non–link-based content (*75*). Future research should examine whether non-link content about vaccines showed different patterns than the ones found in our analysis of URLs.

Another potential limitation is that while our experimental participants were randomly exposed to content, vaccine-hesitant users on Facebook might have actively sought out anti-vaccine content or been selectively targeted to see it by Facebook's algorithm and therefore might be a different population than the one sampled in our experiments. Although we do not find evidence that treatment effects (conditional on exposure) differed significantly on the basis of participant characteristics (including pretreatment 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 SM section S4.7, we analyze the extent to which exposure to hesitancy-inducing 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 hesitancy-inducing (27%). However, all political groups saw at least 10% of such content and, perhaps most concerningly, 23% of content that was viewed by users who do not actively follow political pages [who make up ~75% of the total user base; see the URL Shares documentation for details (*48*)] was predicted to be hesitancy-inducing. Furthermore, the fact that >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.

Another open question pertains to whether exposure to vaccination content in a social media environment might diminish the effects we find in our survey environment. For example, social media users might be less attentive to each story in their newsfeed than our survey participants, who saw a single story rather than a feed of content, or they might discount messages if they are shared by untrustworthy people in their network. However, we found no evidence that less-attentive users were less persuaded by the headlines in our experiments (and, in fact, we found some evidence that less-attentive users exhibited greater levels of persuadability, perhaps because they were less scrutinizing of the evidence quality; see SM section S1.5.6). Furthermore, while source credibility can moderate persuasive effects, other work has shown that even messages from untrustworthy sources can be persuasive (*76*). Future work should further compare how survey results translate to a social media environment.

Despite these limitations, our results have important policy implications, highlighting the need to consider the reach and impact of content—not just its veracity. Whether or not one categorizes content that is misleading without being factually inaccurate as "misinformation" (*19*, *36*, *77*), our findings suggest that this gray area content has the potential to inflict substantial societal harm. Accordingly, 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 that is misleading without being literally false. For example, psychological inoculation is one potential approach that has been shown to help social media users identify "manipulative techniques" that extend beyond outright lying (*78*–*80*). Of equal importance is that mainstream media outlets with widespread reach consider how readers might respond to their reporting in ways that cause real-world harm, despite the caveats and acknowledgment of uncertainty included in their coverage. This is especially relevant in a social media environment where most people only read the headlines and users can present true stories out of context to support misleading narratives (*32*). Rather than focusing exclusively on the accuracy of the facts they report, journalists might 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 potentially misleading but not unambiguously false, 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 trade-off is impossible without being able to quantify the impact of such policies. Here, we provide an approach for conducting such quantification.

Finally, our approach contributes to a growing body of literature offering an alternative to the traditional social science research approach in which a single experiment serves as a test of theory (*81*). Instead, we demonstrate how it is possible to discover which content has an impact from the "bottom-up" rather than relying on the (potentially biased) inclinations of researchers or technology company employees. Our approach offers a replicable framework for researchers (and social media companies) to identify and measure the impact of potentially harmful content in contexts where field experiments are not possible. In contrast to recent work examining the role of Facebook in the 2020 election (*11*, *16*–*18*), which resulted from an intensive, one-off collaboration between academics and Facebook researchers, our method requires only minimal Facebook data access and offers an actionable strategy for continuously improving platforms. Although future work is needed to assess the extent to which this crowd prediction approach generalizes to topics beyond COVID-19 vaccination, we are optimistic that such a method could be replicated by other researchers, both external and internal to Facebook, and applied to new contexts or outcomes (e.g., identifying which content on the social media platform X exacerbates or decreases affective polarization) (*82*–*84*). If shown to be robust, we believe this approach can allow policy-makers to make decisions about mitigating harm that are based on evidence and quantitative assessments, rather than simply on intuition.

Materials and methods summary

For our experimental design, we ran two survey experiments on the online survey platform Lucid (N = 8603 and N = 10,122, respectively) testing the effect of a single exposure to vaccine (mis)information on intentions to take a future COVID-19 vaccine. Participants were exposed to either control or treatment headlines mimicking a social media format and then asked about their intentions to take a future COVID-19 vaccine on a scale of 0 to 100. Then, for our exposure analysis, we used Social Science One and Facebook's URL Shares dataset to identify 13,206 URLs that were related to the COVID-19 vaccine and were popular on Facebook in the US from January to March of 2021. We calculated the number of unique US Facebook users who viewed each URL and labeled content as "flagged misinformation" if it was rated as such by third-party fact-checkers. Finally, we built a crowd-machine pipeline to predict the impact of each Facebook URL on US vaccination intentions. We first recruited 177 laypeople from CloudResearch's Amazon Mechanical Turk panel to predict the persuasive effect of each of the 130 items. We then used a COVID-Twitter-BERT machine-learning model trained on these crowdsourced judgments to predict treatment effects for the entire set of 13,206 URLs from their headlines and descriptions. We explored different models and chose the one with the lowest test set mean squared error (in addition to other evaluation metrics) (*55*). Full details of the experiments, Facebook dataset, and crowd-machine pipeline can be found in the supplementary materials.

REFERENCES AND NOTES

1. D. M. J. Lazer *et al.*, The science of fake news. *Science* **359**, 1094–1096 (2018). doi: 10.1126/science.aao2998; pmid: 29590025
2. C. Shao, G. L. Ciampaglia, A. Flammini, F. Menczer, "Hoaxy: A platform for tracking online misinformation," *WWW '16 Companion: Proceedings of the 25th International Conference Companion on World Wide Web* (International World Wide Web Conferences Steering Committee, 2016), pp. 745–750. doi: 10.1145/2872518.2890098
3. C. Shao *et al.*, Anatomy of an online misinformation network. *PLOS ONE* **13**, e0196087 (2018). doi: 10.1371/journal.pone.0196087; pmid: 29702657
4. S. Tschiatschek, A. Singla, M. Gomez Rodriguez, A. Merchant, A. Krause, "Fake news detection in social networks via crowd signals," *WWW '18: Companion Proceedings of The Web Conference 2018* (International World Wide Web Conferences Steering Committee, 2018), pp. 517–524.
5. S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online. *Science* **359**, 1146–1151 (2018). doi: 10.1126/science.aap9559; pmid: 29590045
6. Y. Wang, M. McKee, A. Torbica, D. Stuckler, Systematic literature review on the spread of health-related misinformation on social media. *Soc. Sci. Med.* **240**, 112552 (2019). doi: 10.1016/j.socscimed.2019.112552; pmid: 31561111
7. J. Allen, B. Howland, M. Mobius, D. Rothschild, D. J. Watts, Evaluating the fake news problem at the scale of the information ecosystem. *Sci. Adv.* **6**, eaay3539 (2020). doi: 10.1126/sciadv.aay3539; pmid: 32284969
8. S. Altay, R. Kleis Nielsen, R. Fletcher, Quantifying the "infodemic": People turned to trustworthy news outlets during the 2020 coronavirus pandemic. *J. Quant. Descr. Digit. Media* **2**, 1–30 (2022). doi: 10.51685/jqd.2022.020
9. N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, D. Lazer, Fake news on Twitter during the 2016 U.S. presidential election. *Science* **363**, 374–378 (2019). doi: 10.1126/science.aau2706; pmid: 30679368
10. A. Guess, J. Nagler, J. Tucker, Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Sci. Adv.* **5**, eaau4586 (2019). doi: 10.1126/sciadv.aau4586; pmid: 30662946
11. S. González-Bailón *et al.*, Asymmetric ideological segregation in exposure to political news on Facebook. *Science* **381**, 392–398 (2023). doi: 10.1126/science.ade7138; pmid: 37499003
12. G. Pennycook, D. G. Rand, Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* **188**, 39–50 (2019). doi: 10.1016/j.cognition.2018.06.011; pmid: 29935897
13. G. Pennycook, D. G. Rand, The psychology of fake news. *Trends Cogn. Sci.* **25**, 388–402 (2021). doi: 10.1016/j.tics.2021.02.007; pmid: 33736957
14. S. van der Linden, A. Leiserowitz, S. Rosenthal, E. Maibach, Inoculating the public against misinformation about climate change. *Glob. Chall.* **1**, 1600008 (2017). doi: 10.1002/gch2.201600008; pmid: 31565263
15. S. van der Linden *et al.*, How can psychological science help counter the spread of fake news? *Span. J. Psychol.* **24**, e25 (2021). doi: 10.1017/SJP.2021.23; pmid: 33840397
16. B. Nyhan *et al.*, Like-minded sources on Facebook are prevalent but not polarizing. *Nature* **620**, 137–144 (2023). doi: 10.1038/s41586-023-06297-w; pmid: 37500978
17. A. M. Guess *et al.*, Reshares on social media amplify political news but do not detectably affect beliefs or opinions. *Science* **381**, 404–408 (2023). doi: 10.1126/science.add8424; pmid: 37499012
18. A. M. Guess *et al.*, How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science* **381**, 398–404 (2023). doi: 10.1126/science.abp9364; pmid: 37498999
19. S. Altay, M. Berriche, A. Acerbi, Misinformation on misinformation: Conceptual and methodological challenges. *Soc. Media Soc.* **9**, 20563051221150412 (2023). doi: 10.1177/20563051221150412
20. S. van der Linden, We need a gold standard for randomised control trials studying misinformation and vaccine hesitancy on social media. *BMJ* **381**, p1007 (2023). doi: 10.1136/bmj.p1007; pmid: 37146997
21. N. Puri, E. A. Coomes, H. Haghbayan, K. Gunaratne, Social media and vaccine hesitancy: New updates for the era of COVID-19 and globalized infectious diseases. *Hum. Vaccin. Immunother.* **16**, 2586–2593 (2020). doi: 10.1080/21645515.2020.1780846; pmid: 32693678
22. J. Aw, J. J. B. Seng, S. S. Y. Seah, L. L. Low, COVID-19 Vaccine Hesitancy-A Scoping Review of Literature in High-Income Countries. *Vaccines* **9**, 900 (2021). doi: 10.3390/vaccines9080900; pmid: 34452026
23. A. Bridgman *et al.*, The causes and consequences of COVID-19 misperceptions: Understanding the role of news and social media. *Harv. Kennedy Sch. Misinformation Rev.* **1**, 10.37016/mr-2020-028 (2020). doi: 10.37016/mr-2020-028
24. F. Pierri *et al.*, Online misinformation is linked to early COVID-19 vaccination hesitancy and refusal. *Sci. Rep.* **12**, 5966 (2022). doi: 10.1038/s41598-022-10070-w; pmid: 35474313
25. S. L. Wilson, C. Wiysonge, Social media and vaccine hesitancy. *BMJ Glob. Health* **5**, e004206 (2020). doi: 10.1136/bmjgh-2020-004206; pmid: 33097547
26. S. Loomba *et al.*, Ability to detect fake news predicts sub-national variation in COVID-19 vaccine uptake across the UK. medRxiv 2023.05.10.23289764 [Preprint] (2023); https://doi.org/10.1101/2023.05.10.23289764.

27. S. Rathje, J. K. He, J. Roozenbeek, J. J. Van Bavel, S. van der Linden, Social media behavior is associated with vaccine hesitancy. *PNAS Nexus* **1**, pgac207 (2022). doi: 10.1093/pnasnexus/pgac207; pmid: 36714849
28. A. A. Arechar *et al*., Understanding and combatting misinformation across 16 countries on six continents. *Nat. Hum. Behav.* **7**, 1502–1513 (2023). doi: 10.1038/s41562-023-01641-6; pmid: 37386111
29. A. M. Guess, B. Nyhan, Z. O'Keeffe, J. Reifler, The sources and correlates of exposure to vaccine-related (mis)information online. *Vaccine* **38**, 7799–7805 (2020). doi: 10.1016/j.vaccine.2020.10.018; pmid: 33164802
30. S. Loomba, A. de Figueiredo, S. J. Piatek, K. de Graaf, H. J. Larson, Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nat. Hum. Behav.* **5**, 337–348 (2021). doi: 10.1038/s41562-021-01056-1; pmid: 33547453
31. C. de Saint Laurent, G. Murphy, K. Hegarty, C. M. Greene, Measuring the effects of misinformation exposure and beliefs on behavioural intentions: A COVID-19 vaccination study. *Cogn. Res. Princ. Implic.* **7**, 87 (2022). doi: 10.1186/s41235-022-00437-y; pmid: 36183027
32. P. Goel, J. Green, D. Lazer, P. Resnik, Mainstream news articles co-shared with fake news buttress misinformation narratives. arXiv:2308.06459 [cs.SI] (2023).
33. J. Benton, "Facebook sent a ton of traffic to a Chicago Tribune story. So why is everyone mad at them?," NiemanLab, 24 August 2021; https://www.niemanlab.org/2021/08/facebook-sent-a-ton-of-traffic-to-a-chicago-tribune-story-so-why-is-everyone-mad-at-them/.
34. L. Wu, F. Morstatter, K. M. Carley, H. Liu, Misinformation in social media: Definition, manipulation, and detection. *SIGKDD Explor.* **21**, 80–90 (2019). doi: 10.1145/3373464.3373475
35. E. C. Tandoc Jr., Z. W. Lim, R. Ling, Defining "fake news": A typology of scholarly definitions. *Digit. Journal. (Abingdon)* **6**, 137–153 (2018). doi: 10.1080/21670811.2017.1360143
36. S. Altay, M. Berriche, H. Heuer, J. Farkas, S. Rathje, A survey of expert views on misinformation: Definitions, determinants, solutions, and future of the field. *Harv. Kennedy Sch. Misinformation Rev.* **4**, 10.37016/mr-2020-119 (2023). doi: 10.37016/mr-2020-119
37. H. Allcott, M. Gentzkow, Social media and fake news in the 2016 election. *J. Econ. Perspect.* **31**, 211–236 (2017). doi: 10.1257/jep.31.2.211
38. Meta, Meta's Third-Party Fact-Checking Program; https://www.facebook.com/formedia/mjp/programs/third-party-fact-checking.
39. O. Ma, B. Feldman, "How Google and YouTube are investing in fact-checking," The Keyword (blog), Google, 29 November 2022; https://blog.google/outreach-initiatives/google-news-initiative/how-google-and-youtube-are-investing-in-fact-checking/.
40. G. Pennycook, D. G. Rand, Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nat. Commun.* **13**, 2333 (2022). doi: 10.1038/s41467-022-30073-5; pmid: 35484277
41. A. D'Ulizia, M. C. Caschera, F. Ferri, P. Grifoni, Fake news detection: A survey of evaluation datasets. *PeerJ Comput. Sci.* **7**, e518 (2021). doi: 10.7717/peerj-cs.518; pmid: 34239967
42. T. Hossain *et al*., "COVIDLies: Detecting COVID-19 misinformation on social media," *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020* (Association for Computational Linguistics, 2020). doi: 10.18653/v1/2020.nlpcovid19-2.11
43. M. Parks, "Few facts, millions of clicks: Fearmongering vaccine stories go viral online," NPR, 25 March 2021; https://www.npr.org/2021/03/25/980035707/lying-through-truth-misleading-facts-fuel-vaccine-misinformation.
44. F. Pierri *et al*., One year of COVID-19 vaccine misinformation on Twitter: longitudinal study. *J. Med. Internet Res.* **25**, e42227 (2023). doi: 10.2196/42227; pmid: 36735835
45. I. J. Borges do Nascimento *et al*., Infodemics and health misinformation: A systematic review of reviews. *Bull. World Health Organ.* **100**, 544–561 (2022). doi: 10.2471/BLT.21.287654; pmid: 36062247
46. E. Chen, J. Jiang, H. H. Chang, G. Muric, E. Ferrara, 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). doi: 10.2196/32378; pmid: 35190798
47. M. Cinelli *et al*., The COVID-19 social media infodemic. *Sci. Rep.* **10**, 16598 (2020). doi: 10.1038/s41598-020-73510-5; pmid: 33024152
48. S. Messing *et al*., Facebook Privacy-Protected Full URLs Data Set, version 10, Harvard Dataverse (2020); https://doi.org/10.7910/DVN/TDOAPG.
49. J. Lasser *et al*., Social media sharing of low-quality news sources by political elites. *PNAS Nexus* **1**, pgac186 (2022). doi: 10.1093/pnasnexus/pgac186; pmid: 36380855
50. A. Guess, K. Aslett, J. Tucker, R. Bonneau, J. Nagler, Cracking open the news feed: Exploring what us Facebook users see and share with large-scale platform data. *J. Quant. Descr. Digit. Media* **1**, 10.51685/jqd.2021.006 (2021). doi: 10.51685/jqd.2021.006
51. S. van der Linden, Misinformation: Susceptibility, spread, and interventions to immunize the public. *Nat. Med.* **28**, 460–467 (2022). doi: 10.1038/s41591-022-01713-6; pmid: 35273402
52. J. Allen, A. A. Arechar, G. Pennycook, D. G. Rand, Scaling up fact-checking using the wisdom of crowds. *Sci. Adv.* **7**, eabf4393 (2021). doi: 10.1126/sciadv.abf4393; pmid: 34516925
53. G. Pennycook, D. G. Rand, Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 2521–2526 (2019). doi: 10.1073/pnas.1806781116; pmid: 30692252
54. J. Kim, B. Tabibian, A. Oh, B. Schölkopf, M. Gomez-Rodriguez, Leveraging the crowd to detect and reduce the spread of fake news and misinformation. arXiv:1711.09918 [cs.SI] (2018).
55. M. Müller, M. Salathé, P. E. Kummervold, COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on Twitter. *Front. Artif. Intell.* **6**, 1023281 (2023). doi: 10.3389/frai.2023.1023281; pmid: 36998290
56. Center for Countering Digital Hate (CCDH), "The disinformation dozen: Why platforms must act on twelve leading online anti-vaxxers" (2022); https://counterhate.com/research/the-disinformation-dozen/.
57. K. Aslett, A. M. Guess, R. Bonneau, J. Nagler, J. A. Tucker, News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions. *Sci. Adv.* **8**, eabl3844 (2022). doi: 10.1126/sciadv.abl3844; pmid: 35522751
58. A. M. Guess *et al*., A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 15536–15545 (2020). doi: 10.1073/pnas.1920498117; pmid: 32571950
59. K. Clayton *et al*., Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Polit. Behav.* **42**, 1073–1095 (2020). doi: 10.1007/s11109-019-09533-0
60. A. Oeldorf-Hirsch, M. Schmierbach, A. Appelman, M. P. Boyle, The ineffectiveness of fact-checking labels on news memes and articles. *Mass Commun. Soc.* **23**, 682–704 (2020). doi: 10.1080/15205436.2020.1733613
61. NewsGuard Tech, NewsGuard–Combating Misinformation with Trust Ratings for News; https://www.newsguardtech.com/.
62. K. Hartwig, F. Doell, C. Reuter, The landscape of user-centered misinformation interventions–a systematic literature review. arXiv:2301.06517 [cs.HC] (2023).
63. P. Patwa *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*, T. Chakraborty, K. Shu, H. R. Bernard, H. Liu, S. Akhtar, Eds., vol. 1402 of Communications in Computer and Information Science (Springer, 2021), pp. 21–29.
64. N. Ruchansky, S. Seo, Y. Liu, "CSI: A hybrid deep model for fake news detection" in *CIKM '17: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (Association for Computing Machinery, 2017), pp. 797–806. doi: 10.1145/3132847.3132877
65. S. Shabani, M. Sokhn, "Hybrid machine-crowd approach for fake news detection" in *2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC)* (The Institute of Electrical and Electronics Engineers, 2018), pp. 299–306.
66. A. Kazemi, K. Garimella, D. Gaffney, S. A. Hale, Claim matching beyond English to scale global fact-checking. arXiv:2106.00853 [cs.CL] (2021).
67. S. Athey, K. Grabarz, M. Luca, N. Wernerfelt, Digital public health interventions at scale: The impact of social media advertising on beliefs and outcomes related to COVID vaccines. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2208110120 (2023). doi: 10.1073/pnas.2208110120; pmid: 36701366
68. R. J. Barro, Vaccination rates and COVID outcomes across U.S. states. *Econ. Hum. Biol.* **47**, 101201 (2022). doi: 10.1016/j.ehb.2022.101201; pmid: 36434953
69. D. A. Patel *et al*., Human papillomavirus vaccine intent and uptake among female college students. *J. Am. Coll. Health* **60**, 151–161 (2012). doi: 10.1080/07448481.2011.580028; pmid: 22316412
70. S. C. Quinn, A. M. Jamison, J. An, G. R. Hancock, V. S. Freimuth, Measuring vaccine hesitancy, confidence, trust and flu vaccine uptake: Results of a national survey of White and African American adults. *Vaccine* **37**, 1168–1173 (2019). doi: 10.1016/j.vaccine.2019.01.033; pmid: 30709722
71. M. H. Danchin *et al*., Vaccine decision-making begins in pregnancy: Correlation between vaccine concerns, intentions and maternal vaccination with subsequent childhood vaccine uptake. *Vaccine* **36**, 6473–6479 (2018). doi: 10.1016/j.vaccine.2017.08.003; pmid: 28811050
72. T. L. Webb, P. Sheeran, Does changing behavioral intentions engender behavior change? A meta-analysis of the experimental evidence. *Psychol. Bull.* **132**, 249–268 (2006). doi: 10.1037/0033-2909.132.2.249; pmid: 16536643
73. P. Sheeran, T. L. Webb, The intention–behavior gap. *Soc. Personal. Psychol. Compass* **10**, 503–518 (2016). doi: 10.1111/spc3.12265
74. G. Godin, M. Conner, Intention-behavior relationship based on epidemiologic indices: An application to physical activity. *Am. J. Health Promot.* **22**, 180–182 (2008). doi: 10.4278/ajhp.22.3.180; pmid: 18251118
75. Y. Yang, T. Davis, M. Hindman, Visual misinformation on Facebook. *J. Commun.* **73**, 316–328 (2023). doi: 10.1093/joc/jqac051
76. B. Clemm von Hohenberg, A. M. Guess, When do sources persuade? The effect of source credibility on opinion change. *J. Exp. Political Sci.* **10**, 328–342 (2023). doi: 10.1017/XPS.2022.2
77. D. Williams, "The fake news about fake news," *Boston Review*, 7 June 2023; https://www.bostonreview.net/articles/the-fake-news-about-fake-news/.
78. C. Lu, B. Hu, Q. Li, C. Bi, X.-D. Ju, Psychological inoculation for credibility assessment, sharing intention, and discernment of misinformation: Systematic review and meta-analysis. *J. Med. Internet Res.* **25**, e49255 (2023). doi: 10.2196/49255; pmid: 37560816
79. J. Roozenbeek, S. van der Linden, B. Goldberg, S. Rathje, S. Lewandowsky, Psychological inoculation improves resilience against misinformation on social media. *Sci. Adv.* **8**, eabo6254 (2022). doi: 10.1126/sciadv.abo6254; pmid: 36001675
80. G. Pennycook *et al*., Misinformation inoculations must be boosted by accuracy prompts to improve judgments of truth. PsyArXiv [Preprint] (2023); https://doi.org/10.31234/osf.io/5a9xq.
81. A. Almaatouq *et al*., Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral sciences. *Behav. Brain Sci.* **47**, e33 (2022). doi: 10.1017/S0140525X22002874; pmid: 36539303
82. C. Timberg, E. Dwoskin, R. Albergotti, "Inside Facebook, Jan. 6 violence fueled anger, regret over missed warning signs," *Washington Post*, 22 October 2021; https://www.washingtonpost.com/technology/2021/10/22/jan-6-capitol-riot-facebook/.
83. J. J. Van Bavel, S. Rathje, E. Harris, C. Robertson, A. Sternisko, How social media shapes polarization. *Trends Cogn. Sci.* **25**, 913–916 (2021). doi: 10.1016/j.tics.2021.07.013; pmid: 34429255
84. E. Kubin, C. von Sikorski, The role of (social) media in political polarization: A systematic review. *Ann. Int. Commun. Assoc.* **45**, 188–206 (2021). doi: 10.1080/23808985.2021.1976070
85. Vaccine Misinformation Project, OSF (2024); https://doi.org/10.17605/OSF.IO/68MN9.

ACKNOWLEDGMENTS

We thank A. Bear, H. Chen, D. Eckles, Z. Epstein, G. Pennycook, L. Hewitt, N. Stagnaro, and B. Tappin for providing valuable feedback on early drafts of this work. **Funding:** We gratefully acknowledge Alain Rossmann for his generous funding of this work. In addition, we thank Richard Jay Mack for funding support for D.J.W. We also acknowledge funding from the MIT Sloan Health Systems Initiative. **Author contributions:** J.A. developed the research concept, designed the study, conducted the study, and analyzed the data under the supervision of D.G.R. D.J.W. secured data access to the URL Shares dataset. J.A. created the figures and wrote the initial draft of the paper, with input from D.G.R. and D.J.W. All authors contributed to reviewing and editing of the manuscript. **Competing interests:** Other work of J.A. has been funded by Meta. Other work of D.G.R. has been funded by Meta and Google. D.J.W. has no competing interests to declare. **Data and materials availability:** Code and data to reproduce the findings in this manuscript can be found in OSF (*85*). Access to the full Meta URL Shares dataset can be requested here: https://socialscience.one/rfps (*48*).

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adk3451
Materials and Methods
Supplementary Text
Figs. S1 to S19
Tables S1 to S35
References (*86*–*106*)
MDAR Reproducibility Checklist

Submitted 29 August 2023; accepted 17 April 2024
10.1126/science.adk3451