CORONAVIRUS

The efficacy of Facebook's vaccine misinformation policies and architecture during the COVID-19 pandemic

David A. Broniatowski[1,2]*, Joseph R. Simons[3], Jiayan Gu[4], Amelia M. Jamison[5], Lorien C. Abroms[2,4]

Online misinformation promotes distrust in science, undermines public health, and may drive civil unrest. During the coronavirus disease 2019 pandemic, Facebook—the world's largest social media company—began to remove vaccine misinformation as a matter of policy. We evaluated the efficacy of these policies using a comparative interrupted time-series design. We found that Facebook removed some antivaccine content, but we did not observe decreases in overall engagement with antivaccine content. Provaccine content was also removed, and antivaccine content became more misinformative, more politically polarized, and more likely to be seen in users' newsfeeds. We explain these findings as a consequence of Facebook's system architecture, which provides substantial flexibility to motivated users who wish to disseminate misinformation through multiple channels. Facebook's architecture may therefore afford antivaccine content producers several means to circumvent the intent of misinformation removal policies.

INTRODUCTION

Online misinformation undermines trust in scientific evidence (*1*) and medical recommendations (*2*). It has been linked to harmful offline behaviors including stalled public health efforts (*3*), civil unrest (*4*), and mass violence (*5*). The coronavirus disease 2019 (COVID-19) pandemic has spurred widespread concern that misinformation spread on social media may have lowered vaccine uptake rates (*6*, *7*). Therefore, policymakers and public officials have put substantial pressure on social media platforms to curtail misinformation spread (*8*, *9*).

Years of "soft" remedies—such as warning labels by Twitter (*10*), YouTube (*11*), and Facebook (*12*) and attempts by these platforms to downrank objectionable content in search—have demonstrated some success (*13*); however, misinformation continues to spread widely online, leading many to question the efficacy of these interventions (*14*). Some have suggested that combining these soft remedies with "hard" remedies (*15*)—removing content and objectionable accounts (*16*–*18*)—could largely curtail misinformation spread (*19*). However, evidence for the short-term efficacy of hard remedies is mixed (*20*–*24*), and the long-term efficacy of these strategies has not been systematically examined. Hard remedies have also spurred accusations of censorship and threats of legal action (*25*, *26*). There is therefore a critical need to understand whether this combination of remedies is effective—i.e., whether it reduces users' exposure to misinformation—and if not, why not.

Any evaluation of the efficacy of these remedies must be grounded in a scientific understanding of why misinformation spreads online. Prior work indicates that misinformation may spread widely on social media if it is framed in a manner that is more compelling than true information (*27*, *28*). Users appear to prefer sharing true information when cued to think about accuracy (*13*); however, the social media environment may interfere with peoples' ability to distinguish truth from falsehood (*29*, *30*). Other studies have suggested that social media platforms' algorithms facilitate the creation of "echo chambers" (*31*, *32*), which increase exposure to content from like-minded individuals. Accordingly, prior interventions have focused on altering the social media environment to either reduce users' exposure to misinformation or inform them when content is false. However, recent evidence suggests that people use online algorithms to actively seek out and engage with misinformation (*33*). Therefore, on the basis of prior theory (*34*, *35*), we examine how a social media platform's "system architecture"—its designed structure that shapes how information flows through the platform (*34*)—enables antivaccine content producers and users to flexibly (*36*) establish new paths to interdicted content, facilitating resistance to content removal efforts.

We analyzed Facebook because it is the world's largest social media platform. In December 2020, when COVID-19 vaccines first became available, Facebook had 2.80 billion monthly active users (*37*). As of April 2023, this number had grown to 2.99 billion monthly active users (*25*). We therefore conducted an evaluation of Facebook's attempts to remove antivaccine misinformation from its public content throughout the COVID-19 pandemic.

Of primary interest were the following three research questions: (i) Were Facebook's policies associated with a substantial decrease in public antivaccine content and engagement with remaining antivaccine content? (ii) Did misinformation decrease when these policies were implemented? (iii) How might Facebook's system architecture have enabled or undermined these policies?

Data and definitions

We downloaded public data from Facebook pages and groups using CrowdTangle (*38*), which has been called "the most effective transparency tool in the history of social media" (*39*). To obtain these data, we searched CrowdTangle on 15 November 2020, immediately

[1]Department of Engineering Management and Systems Engineering, The George Washington University, Washington, DC 20052, USA. [2]Institute for Data, Democracy, and Politics, The George Washington University, Washington, DC 20052, USA. [3]Office of the Assistant Secretary for Financial Resources, United States Department of Health and Human Services, Washington, DC 20543, USA. [4]Department of Prevention and Community Health, The George Washington University, Washington, DC 20052, USA. [5]Department of Health, Behavior, and Society, Johns Hopkins University, Baltimore, MD 21218, USA.
*Corresponding author. Email: broniatowski@gwu.edu

before Facebook's 18 November 2020 removal of "Stop Mandatory Vaccination"—one of the platform's largest antivaccine fan pages (*40*). We chose this event because it immediately preceded Facebook's 3 December 2020 announcement that false claims about COVID-19 vaccines and accounts that repeatedly posted these claims would be removed (*41*). Furthermore, on 8 February 2021, Facebook extended this policy to vaccine misinformation in general (*42*). Thus, our reference date marks the first of a series of hard content remedies specifically targeting vaccine misinformation.

Our search yielded a set of 216 (114 antivaccine and 102 provaccine) English-language pages and 100 (92 antivaccine and 8 provaccine) English-language groups that primarily discussed vaccines. Our dataset consisted of 119,091 (86,893 antivaccine, 73%) posts to pages and 168,419 (163,095 antivaccine, 97%) posts to groups that were created between 15 November 2019 and 15 November 2020 and 177,615 (110,093 antivaccine, 62%) and 244,981 (231,788 antivaccine, 95%) posts that were created between 16 November 2020 and 28 February 2022 to the same pages and groups, respectively.

We examined pages and groups separately because they serve different functions in Facebook's system architecture: Only page administrators may post in pages, which are designed for marketing and brand promotion. In contrast, any member may post in groups, which serve as a forum for members to build community and discuss shared interests. Furthermore, pages may serve as group administrators (*43*), establishing an architecturally meaningful hierarchical relationship between these two types of venues.

We used a comparative interrupted time-series (CITS) design to compare the weekly number of Facebook posts in public antivaccine and provaccine Facebook pages and groups to prepolicy trends and to compare these trends to one another. Beyond raw post volumes, we measured the total number of engagements—defined as sum of all interactions (shares, comments, likes, and emotional reactions, such as sad, love, angry, etc.)—with each post. Facebook uses these engagements when prioritizing content in users' newsfeeds (*44*, *45*), meaning that content with many engagements is more likely to be seen and, consequently, shared. We therefore consider Facebook's policies to be efficacious if they reduced engagement with antivaccine content compared to prepolicy trends.

RESULTS

Forty-nine (37, 76% antivaccine) pages and 31 (28, 90% antivaccine) groups in our sample were removed by 28 February 2022. Antivaccine pages and groups were 2.13 [95% confidence interval (CI), 1.70 to 2.66] times more likely to have been removed than their provaccine counterparts. In addition, 5 (5%) antivaccine groups changed their settings from public to private.

Antivaccine and provaccine content volumes decreased

Posts in both antivaccine and provaccine pages decreased relative to prepolicy trends (Fig. 1A). However, antivaccine page post volumes decreased more than provaccine page post volumes: Antivaccine post volumes decreased 1.47 [relative risk (RR) = 0.68; 95% CI, 0.61 to 0.76; Table 1] times more than provaccine post volumes. Posts to antivaccine groups also decreased relative to prepolicy trends (Fig. 1B), and antivaccine group post volumes decreased 3.57 (RR = 0.28; 95% CI, 0.21 to 0.37; Table 1) times more than provaccine group post volumes.

Engagement with antivaccine content matched or exceeded prepolicy expectations

We did not observe significant changes in engagement with content in antivaccine pages (RR = 0.73; 95% CI, 0.28 to 1.90; Fig. 1E). In groups, antivaccine engagement counts were, on average, 33% higher than what would be expected on the basis of prepolicy trends (RR = 1.33; 95% CI, 1.05 to 1.69; Fig. 1F), although this increase was not significant when compared to provaccine group trends (RR = 1.22; 95% CI, 0.94 to 1.56; Table 1). In addition, while performing a robustness check on a second sample of antivaccine groups that were identified on 28 July 2021, we found that engagement counts were comparable to prepolicy levels (fig. S10), raising the possibly that Facebook's policies may have shifted, rather than decreased, engagement.

Misinformative topics increased in antivaccine pages

We next examined whether remaining content became less misinformative following Facebook's policies. We observed increases in the proportions of several topics appearing to violate Facebook's community standards in Facebook pages (Fig. 2; topics are shown in table S1, and examples posts are shown in table S2). The largest increase occurred in a topic alleging severe COVID-19 vaccine adverse reactions [odds ratio (OR) = 1.41; 95% CI, 1.05 to 1.90; table S3]. Similarly, reports of hospitalization and death (OR = 1.23; 95% CI, 1.80 to 1.38) and promotion of alternative medicine (OR = 1.32; 95% CI, 1.28 to 1.35) increased. Topics alleging negative effects of vaccines on immunity, either due to toxic ingredients (OR = 1.24; 95% CI, 1.13 to 1.37) or focused on children (OR = 1.34; 95% CI, 1.02 to 1.77), also increased. We also observed increases in topics discussing school (OR = 2.02; 95% CI, 1.65 to 2.46) and other vaccine mandates (OR = 1.21; 95% CI, 1.16 to 1.27), legislation opposing vaccination (OR = 1.06; 95% CI, 1.02 to 1.13), and antivaccine medical advice (OR = 1.14; 95% CI, 1.09 to 1.18).

Discussion of several misinformative topics also increased in antivaccine groups, compared to prepolicy trends; especially those pertaining to aspects of vaccine safety and efficacy, as well as mandates (table S3). These increases were significantly smaller in magnitude than increases in provaccine groups, suggesting that they were not attributable to Facebook's policies.

Links to misinformative and polarized sources increased in antivaccine groups

We next examined whether the proportion of links to low credibility (*46*, *47*) external sources changed. Beyond spreading potential misinformation, these links have architectural importance because they enable users to easily access content off the platform that is not under the control of Facebook's content moderators. In pages, we observed an increase in low-credibility links between February and mid-September 2021; however, the proportion of these links subsequently decreased, leading to an overall null effect (OR = 1.00; 95% CI, 0.96 to 1.04; Fig. 3A and Table 1). A whistleblower at Facebook released several internal documents to the Securities and Exchange Commission and the Wall Street Journal (*48*) in mid-September 2021, garnering substantial media attention. This media attention may have led Facebook to target these links for removal; however, this is a post hoc speculation. Despite this decrease, the odds that a

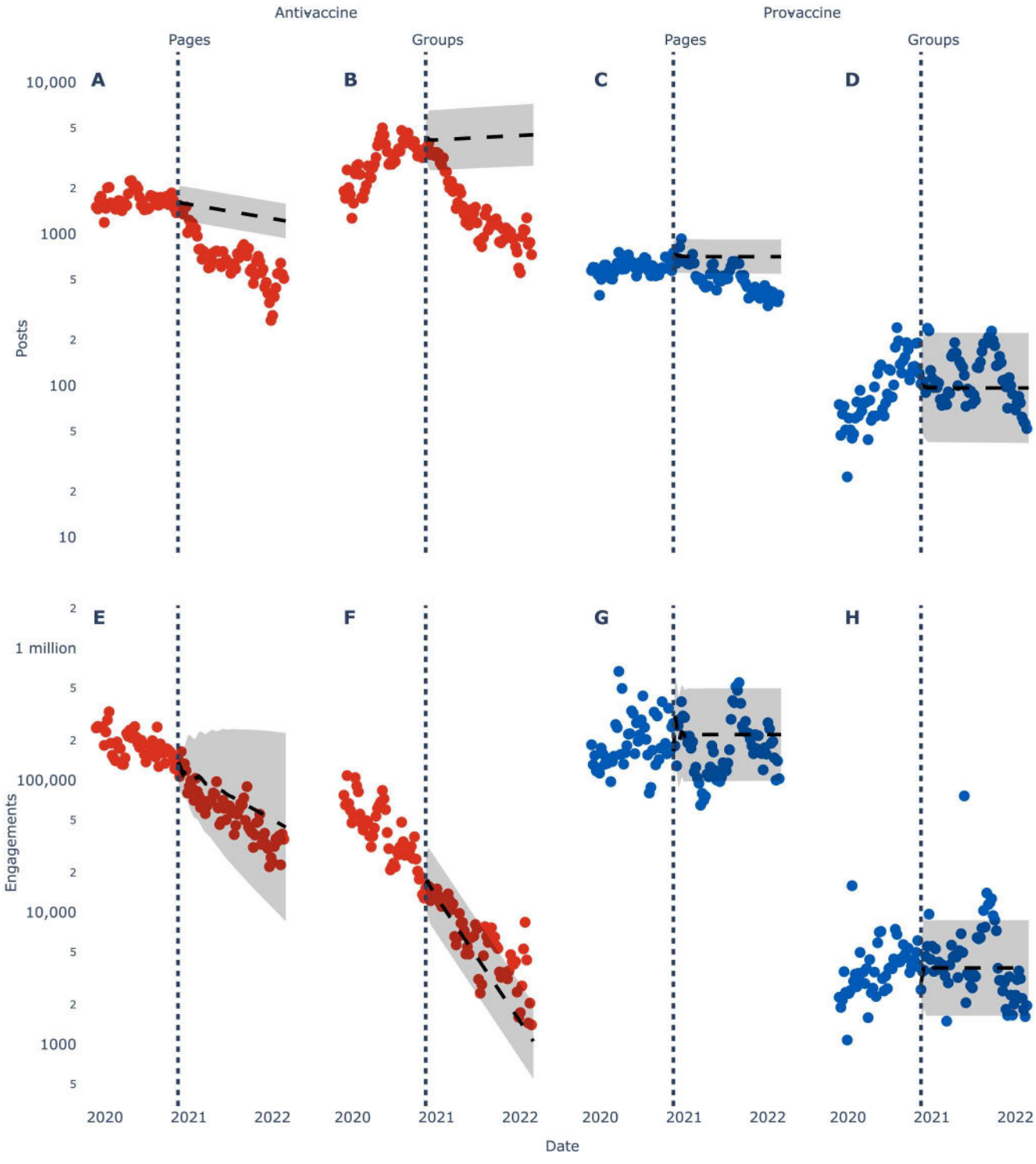

Fig. 1. Comparison of empirical post and engagement counts to prepolicy trends. Projections based on autoregressive integrated moving average (ARIMA) models are projected forward from the end of the prepolicy period (15 November 2020; dotted line). Analyses were conducted after applying a logarithmic transform to the data; therefore, all data are displayed on a logarithmic scale. Gray indicates 95% CIs. Antivaccine content and engagements are shown in red, and provaccine content and engagements are shown in blue. (**A**) Posts in antivaccine pages were significantly below prepolicy trends. (**B**) Posts in antivaccine groups were significantly below prepolicy trends. (**C**) Posts in provaccine groups were significantly below prepolicy trends. (**D**) Posts in provaccine groups were consistent with prepolicy trends. (**E**) Engagements with posts in antivaccine pages were consistent with prepolicy trends. (**F**) Engagements with posts in antivaccine groups significantly exceeded prepolicy trends. (**G**) Engagements with posts in provaccine pages were consistent with prepolicy trends. (**H**) Engagements with posts in provaccine groups were consistent with prepolicy trends.

Table 1. Results of CITS analyses. OR, odds ratio; N/A, not applicable (e.g., because misinformative or polarized links were rare in provaccine pages and groups).

Venue type	Count type	Stance	Data compared to prepolicy trend RR (95% CI)	*P*	Change in antivaccine compared to change in provaccine RR (95% CI)	*P*
			Counts			
Pages	Posts	Anti	0.49 (0.45–0.53)	<0.001	0.68 (0.61–0.76)	<0.001
		Pro	0.72 (0.67–0.76)	<0.001		
	Engagements	Anti	0.73 (0.28–1.90)	0.52	0.94 (0.36–2.45)	0.89
		Pro	0.78 (0.69–0.88)	<0.001		
Groups	Posts	Anti	0.32 (0.27–0.38)	<0.001	0.28 (0.21–0.37)	<0.001
		Pro	1.15 (0.92–1.45)	0.21		
	Engagements	Anti	1.33 (1.05–1.69)	0.02	1.22 (0.94–1.56)	0.14
		Pro	1.09 (0.97–1.23)	0.13		
			Data compared to prepolicy trend OR (95% CI)	*P*	Change in antivaccine compared to change in provaccine OR (95% CI)	*P*
			Iffy URLs			
Pages	Posts	Anti	1.00 (0.96–1.04)	0.97	N/A	N/A
	Engagements	Anti	2.69 (2.22–3.27)	<0.001	N/A	N/A
Groups	Posts	Anti	1.21 (1.02–1.42)	0.03	N/A	N/A
	Engagements	Anti	1.07 (0.49–2.31)	0.87	N/A	N/A
			Polarized URLs			
Pages	Posts	Anti	1.35 (1.24–1.48)	<0.001	1.30 (1.09–1.54)	0.003
		Pro	1.04 (0.90–1.21)	0.57		
	Engagements	Anti	2.37 (1.03–5.45)	0.04	N/A	N/A
Groups	Posts	Anti	1.51 (1.38–1.66)	<0.001	N/A	N/A
	Engagements	Anti	0.92 (0.85–1.00)	0.06	N/A	N/A
			Links to other Facebook pages			
Pages	Posts	Anti	0.30 (0.25–0.36)	<0.001	0.29 (0.03–3.62)	0.31
		Pro	1.01 (0.10–10.47)	0.99		
	Engagements	Anti	0.10 (0.08–0.14)	<0.001	0.17 (0.11–0.26)	<0.001
		Pro	0.61 (0.46–0.81)	<0.001		
			Coordinated pages and groups			
Combined	Posts	Anti	0.41 (0.37–0.46)	<0.001	N/A	N/A
			Reaction proportions			
Pages	Angry	Anti	0.49 (0.47–0.52)	<0.001	0.10 (0.02–0.64)	0.01
		Pro	4.80 (0.77–29.90)	0.09		
	Other reactions	Anti	1.43 (1.39–1.48)	<0.001	2.71 (0.20–36.62)	0.44
		Pro	0.53 (0.04–7.15)	0.63		
Groups	Angry	Anti	0.54 (0.49–0.59)	<0.001	0.30 (0.25–0.37)	<0.001
		Pro	1.77 (1.49–2.11)	<0.001		
	Other reactions	Anti	1.30 (1.24–1.37)	<0.001	1.18 (1.12–1.24)	<0.001
		Pro	1.10 (1.08–1.12)	<0.001		
			Absolute change from prepolicy trend (95% CI)	*P*	Relative change in antivaccine compared to provaccine (95% CI)	*P*
			Partisan bias			
Pages	Posts	Anti	0.28 (0.20–0.36)	<0.001	0.37 (0.26–0.47)	<0.001
		Pro	−0.09 (−0.16 - −0.02)	0.02		
	Engagements	Anti	0.01 (−0.02–0.03)	0.65	0.03 (−0.04–0.09)	0.47

continued on next page

Venue type	Count type	Stance	Data compared to prepolicy trend RR (95% CI)	P	Change in antivaccine compared to change in provaccine RR (95% CI)	P
		Pro	-0.02 (−0.08–0.04)	0.55		
Groups	Posts	Anti	0.24 (−0.15–0.64)	0.24	−0.54(−2.79–1.70)	0.63
		Pro	0.78 (−1.42–2.99)	0.48		
	Engagements	Anti	0.71 (0.24–1.18)	0.003	0.90 (0.40–1.40)	<0.001
		Pro	−0.19 (−0.36–0.02)	0.03		

user engaged with posts containing these links was 2.69 (95% CI, 2.22 to 3.27) times higher than expectations based on prepolicy trends (Fig. 3B). In groups, the odds that a post contained a link to a low credibility source was 1.21 (95% CI, 1.02 to 1.42) times higher than expectations based on prepolicy trends (Fig. 3C).

These links may have increasingly exposed antivaccine audiences to politically polarized content. Relative to prepolicy trends, the odds that a post contained a link to a politically polarized website increased in pages (OR = 1.35; 95% CI, 1.24 to 1.48; Fig. 3E) and groups (OR = 1.51; 95% CI, 1.38 to 1.66; Fig. 3G). Engagement with these links increased in antivaccine pages (OR = 2.37; 95% CI, 1.03 to 5.45; Fig. 3H).

On a nine-point scale, the average link posted in antivaccine pages that was rated for political stance became 0.28 (95% CI, 0.20 to 0.36) points more polarized toward the right wing compared to prepolicy trends. In contrast, politically rated links became 0.09 (95% CI, 0.02 to 0.16) points more polarized toward the left wing in provaccine pages, leading to an overall widening of the partisan gap by 0.37 (95% CI, 0.26 to 0.47) points (Fig. 3I). A trend toward rated links in antivaccine groups displaying increasing right-wing partisan bias throughout the prepolicy period continued through February 2022 (Fig. 3K). Last, the average engagement with a rated link in antivaccine and provaccine groups became 0.71 (95% CI, 0.24 to 1.18) and 0.19 (95% CI, 0.02 to 0.36) points more polarized to the political right and left, respectively, leading to an overall widening of the partisan gap by 0.90 (95% CI, 0.40 to 1.40) points (Fig. 3L).

Motivated users may still seek out misinformation

Facebook's attempts to remove vaccine misinformation may have faltered if the desire to engage with misinformation remained high, despite Facebook's attempts to remove content. Several converging lines of evidence support this hypothesis. First, prior work suggests that removal of antivaccine Facebook groups may have been associated with increased activity on other platforms, such as Twitter (*49*). Second, we observed an increase in links to "alternative" social media platforms—BitChute, Rumble, and Gab (*50*, *51*)—as YouTube and Twitter began removing COVID-19 misinformation on 14 October 2020 and 1 March 2021, respectively. Links to these alternative platforms were more prevalent in both pages, RR = 2.88 (95% CI, 2.38 to 3.48), $P < 0.001$, and groups, RR = 3.41 (95% CI, 3.19 to 3.64), $P < 0.001$, after the prepolicy period. Furthermore, posts containing these links had a larger share of engagements in both pages, RR = 1.60 (95% CI, 1.55 to 1.65), $P < 0.001$, and groups, RR = 7.19 (95% CI, 6.94 to 7.45), $P < 0.001$. Third, a simulation model that we constructed and calibrated to prepolicy data (see the Supplementary Materials), best reproduced data from after the prepolicy period when we assumed that removals shifted demand for antivaccine content to remaining pages and groups, rather than reducing it. Together, these results suggest that content and account removals may not have dissuaded audiences from seeking out antivaccine misinformation.

Facebook's architecture allows them to find it

Even if demand for antivaccine content remains high, Facebook users require the ability to find and access it, despite removals, for this explanation to be plausible. Flexible systems enable users to access interdicted content via several alternative paths, even if some paths are removed (*34*–*36*). We observe that Facebook has three key design features that, together, constitute a "layered hierarchy" (Fig. 4)—a structure that is theorized to promote flexibility in the face of removals, while simultaneously allowing antivaccine page administrators some degree of control over information flow (*34*, *36*). This combination of flexibility and control arises from the ease with which page administrators and users may establish new paths to interdicted content, while still respecting a hierarchical structure in which pages can administer groups (*43*) and in which content posted in both pages and groups is likely to be highly ranked in users' newsfeeds. Antivaccine page administrators can use this structure to disseminate content to users in a manner that is mediated by participatory (*52*) discussion groups that facilitate awareness of and demand for antivaccine content. We therefore hypothesize that each layer in this hierarchy facilitates flexibility by enabling users to easily access antivaccine content despite removals: (i) as followers of pages administered by antivaccine opinion leaders, (ii) as members of discussion groups that may coordinate to post redundant content, and (iii) by interacting with (e.g., liking, sharing, etc.) posts to promote them in users' newsfeeds.

Top layer: Pages are still reachable from one another

The top layer of Facebook's hierarchy consists of pages that connect to each other. Page administrators can collaborate to share followers and content by linking to, or "liking," (*53*) one another's pages. Users can then follow these links, finding new pages and content. Thus, linked pages likely have overlapping sets of followers, creating flexibility. Consistent with this explanation, antivaccine and provaccine pages frequently posted links to other aligned pages, forming densely connected clusters (Fig. 4A) that predated Facebook's policies and likely share audiences.

A system is flexible if it allows users to easily make changes to the system without altering its overall architecture (*34*, *35*). Here, flexibility means that users have several means to easily access antivaccine content even if some ways of accessing that content was removed (e.g., via post or account deletion). Because Facebook uses a flexible layered hierarchy, removing some pages may not

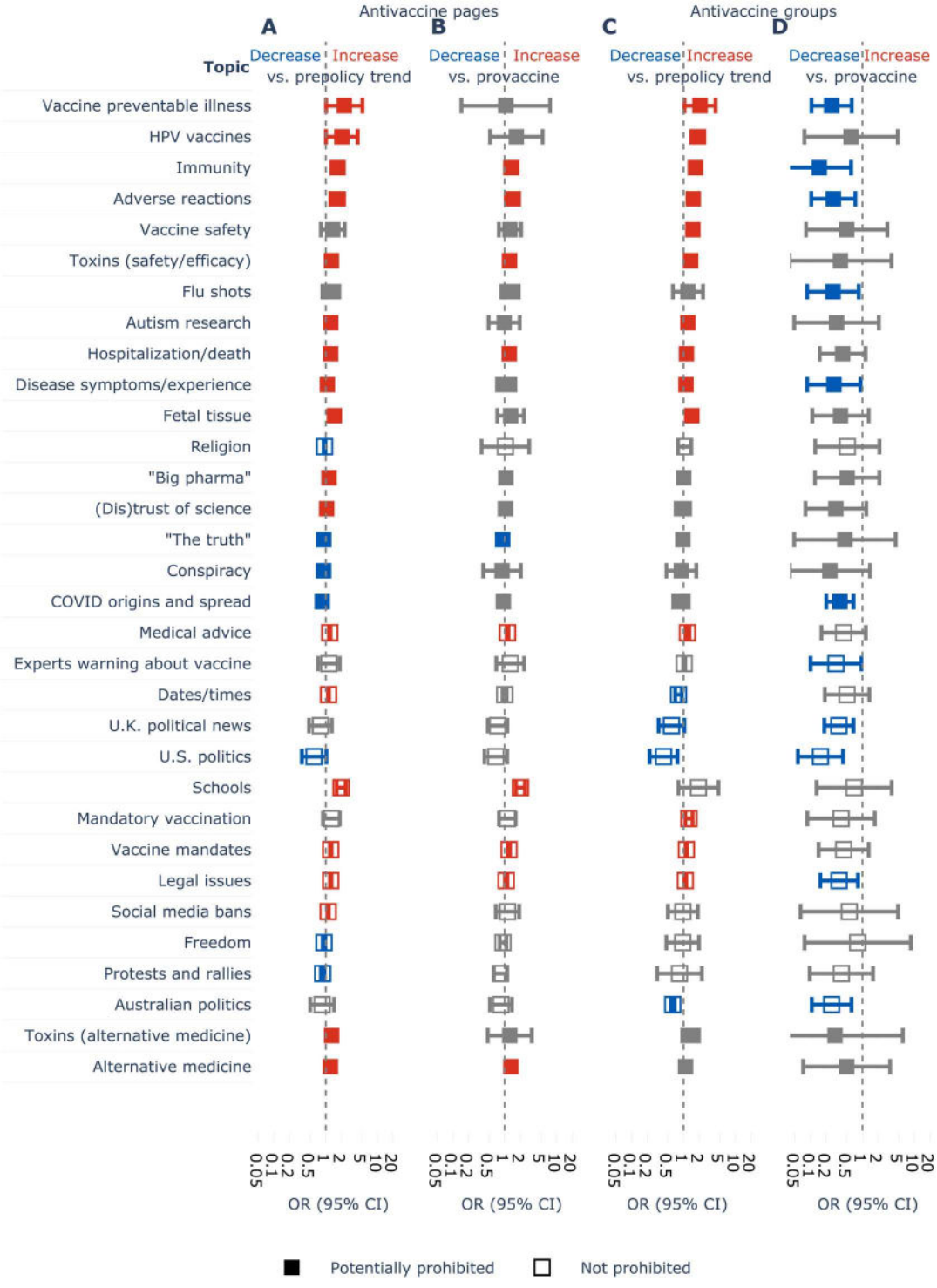

Fig. 2. Changes in antivaccine topic proportions relative to prepolicy and provaccine trends. A dashed line indicates an OR of 1, meaning no change. Topics are shown if at least one comparison was statistically significant. Ninety-five percent of CIs were calculated after applying the Benjamini-Hochberg procedure (*77*) to correct for multiple comparisons, with a false discovery rate of α = 0.05. Markers are coded according to a validated typology of online vaccine content (*64*, *65*). Solid markers indicate that the topic is potentially prohibited under Facebook's policies (*76*). Red (blue) markers indicate topics that increased (decreased) significantly relative to either prepolicy or provaccine trends. (**A**) In pages, topics related to vaccine safety and efficacy increased in proportion relative to prepolicy trends, whereas those relative to conspiracy theories decreased. (**B**) In pages, topics related to vaccine safety and efficacy increased relative to changes in provaccine topics. (**C**) In groups, topics related to vaccine safety and efficacy increased in proportion relative to prepolicy trends, whereas those relative to politics decreased. (**D**) Several topics decreased in groups relative to larger increases in provaccine groups. HPV, human papillomavirus.

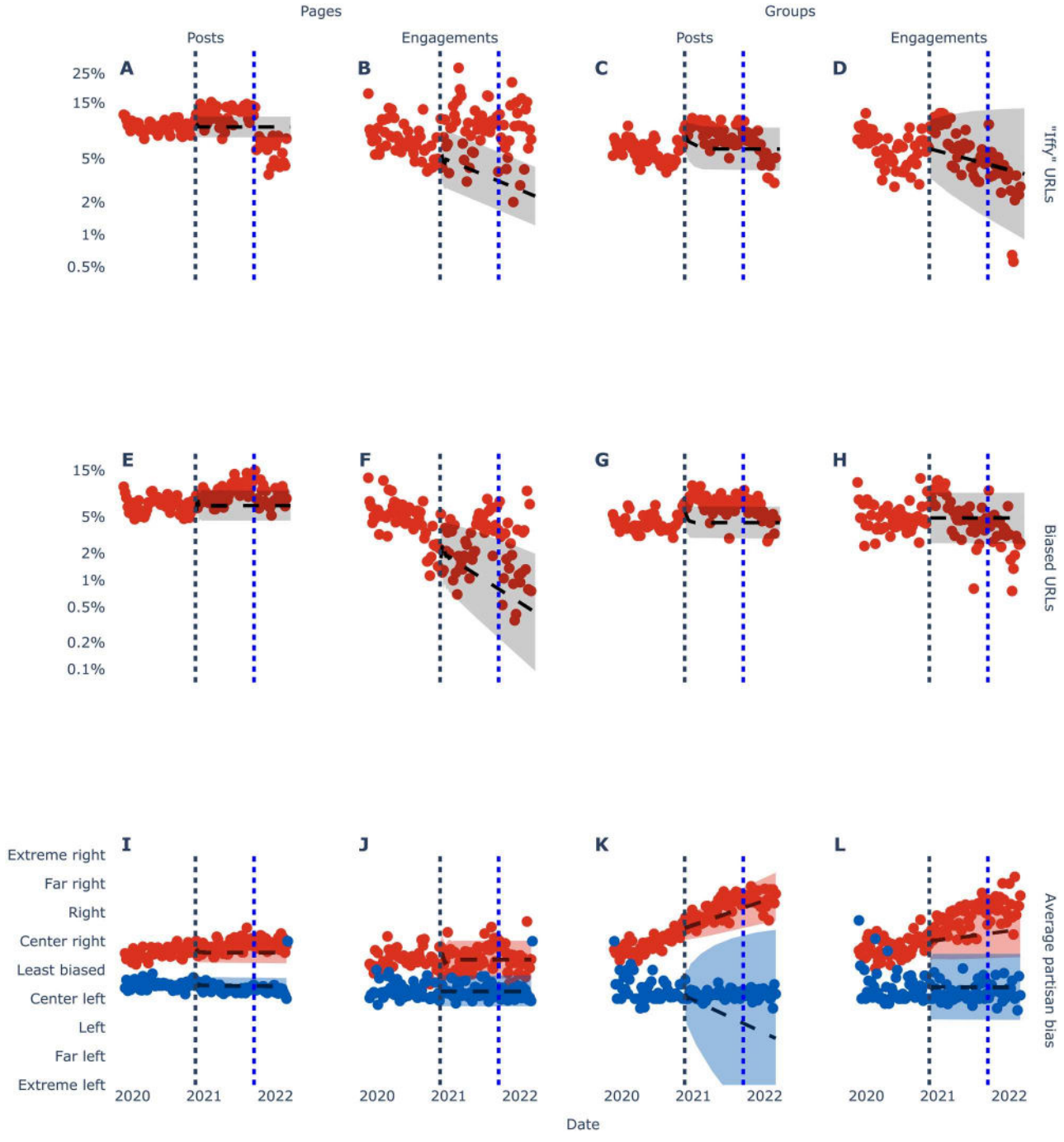

Fig. 3. Comparison of URL source credibility and partisanship ratings to prepolicy data. Dashed lines represent ARIMA model projections from the end of the prepolicy period (15 November 2020, black dotted line). On 14 September 2021 (blue dotted line), a whistleblower at Facebook released internal documents to the Wall Street Journal and Securities and Exchange Commission. The data in the top two rows are displayed on a logit scale due to logistic transformation during analysis. Antivaccine and provaccine content and engagements are depicted in red and blue, respectively. Ninety-five percent of CIs are shown. (**A**) Proportions of Iffy URLs in antivaccine pages exceeded prepolicy trends until 14 September 2021 and subsequently dropped below those trends. (**B**) The proportion of engagements with Iffy URLs in antivaccine pages exceeded prepolicy trends. (**C**) Proportions of Iffy URLs in antivaccine groups exceeded prepolicy trends. (**D**) The proportion of engagements with Iffy URLs in antivaccine groups remained consistent with prepolicy trends. (**E**) Proportions of polarized URLs in antivaccine pages exceeded prepolicy trends. (**F**) The proportion of engagements with polarized URLs in antivaccine pages exceeded prepolicy trends. (**G**) Proportions of polarized URLs in antivaccine groups exceeded prepolicy trends. (**H**) The proportion of engagements with polarized URLs in antivaccine groups remained consistent with prepolicy trends. (**I**) On average, URLs rated for political stance in anti- and provaccine pages became more right- and left-leaning, respectively. (**J**) When weighted by engagement, the average ratings among URLs rated for political stance in both antivaccine and provaccine pages remained consistent with prepolicy trends. (**K**) On average, URLs rated for political stance in both antivaccine and provaccine groups remained consistent with prepolicy trends, indicating a sustained rightward shift in antivaccine groups. (**L**) When weighted by engagement numbers, the average political stance rating became more right- and left-leaning in antivaccine and provaccine groups, respectively.

Fig. 4. Pages, groups, and newsfeeds form a layered hierarchy that facilitates access to, and demand for, antivaccine content despite hard remedies. In the top layer, page administrators link to one another's pages to create overlapping audiences. Groups form the middle layer of Facebook's layered hierarchy because pages can serve as group administrators (*43*). Group members and page administrators can expand the audience for their content by posting it in multiple venues simultaneously. The top two layers are network diagrams with nodes representing Facebook pages (circles) or groups (squares). Edges that were present during the prepolicy period only are in gray; edges formed after 15 November 2020 are in fuchsia; edges that were present before and after 15 November 2020 are in black. Node locations are determined using a force-directed layout (*88*). In the bottom layer, content enters individual users' newsfeeds according to the accounts that they follow. Content spurring more reactions is more highly ranked in these newsfeeds, and thus more likely to be seen. In this layer, "x"-shaped markers represent the proportion of angry reactions relative to all engagements (reactions, comments, and shares). Triangular markers represent the all other reactions (like, love, haha, wow, sad, and care) relative to all engagements. Projections based on ARIMA models are projected forward from the end of the prepolicy period (15 November 2020; dotted line). Analyses were conducted after applying a logistic transform to the data; therefore, data are displayed on a logit scale. Ninety-five percent of CIs are shown. (**A**) Edges indicate a URL pointing from one page to another. (**B**) Edges indicate that these venues routinely post the same URL near simultaneously (within 33 s). In both antivaccine pages (**C**) and antivaccine groups (**D**), the proportion of angry reactions decreased, and the proportion of all other reactions increased relative to prepolicy trends.

have prevented access to remaining pages for many followers. We examined whether Facebook's policies might have eroded the flexibility of the antivaccine cluster by removing enough paths through it to separate it into multiple components, thus making some pages unreachable from other pages. Although the odds that a post contained a link to another antivaccine page decreased to 30% (95% CI, 0.25 to 0.36; fig. S1) of prepolicy trends, this reduction was neither significant when compared to changes within the provaccine cluster (OR = 0.29; 95% CI, 0.03 to 3.62; Table 1) nor sufficient to separate the cluster into multiple components (Fig. 4A and fig. S2). Thus, we cannot conclude that Facebook's policies meaningfully reduced the flexibility of the antivaccine cluster.

Middle layer: New antivaccine groups coordinate with older groups

The middle layer of Facebook's hierarchy consists of groups and pages that coordinate with each other. This creates a second source of flexibility because users can reach larger audiences by sharing the same content in multiple venues simultaneously. Groups and pages that coordinate in this manner can circumvent Facebook's content removal policies by generating posts with the same content such that, if one post is removed, copies of it will still be present on the platform. Removal of individual groups and pages will be less effective if users post the same content elsewhere.

Before the policy, we observed that several antivaccine pages and groups routinely posted the same content "near simultaneously"—a sign of potential "coordinated inauthentic behavior," which is explicitly prohibited by Facebook (*54*). After the policy, several of these pages and groups continued to do so, coordinating with new groups (Fig. 4B). Corresponding links often promoted antivaccine Facebook pages, content on other social media platforms (e.g., YouTube), known antivaccine websites (e.g., ageofautism.com), and websites promoting political calls to action, such as petitions opposing mandatory vaccines. In contrast, coordinated provaccine content pointed to websites facilitating COVID vaccination, such as vaccinefinder.org.

We examined whether the proportion of these "near-simultaneous" link pairs decreased after Facebook's policy. Coordination in provaccine venues was absent for several months before 16 August 2020. We therefore only fit trends to antivaccine data, finding a significant decrease in the odds that a pair of links were coordinated between groups and pages, to 41% of the prepolicy trend (95% CI, 0.37 to 0.46; Table 1 and fig. S3). Facebook reduced visible signs of coordination in the antivaccine cluster, by breaking it into multiple, disconnected components; however, new groups and links also appear to have formed, reconnecting these disconnected clusters, which continued to engage in coordinated sharing (Fig. 4B and fig. S4). Briefly, groups and pages were able to establish new connections when old ones were disrupted, restoring lost flexibility (*36*).

Bottom layer: Antivaccine posts are increasingly likely to be promoted in newsfeeds

The bottom layer of Facebook's hierarchy is made up of all users who might see antivaccine content in their newsfeeds. Here, a third source of flexibility makes use of Facebook's newsfeed algorithm, which is designed to promote content that has generated "meaningful social interaction" (*45*, *55*). Facebook reportedly uses a weighted sum of reactions (e.g., "like," "love," "angry," etc.) that content has spurred to rank content in users' newsfeeds (Fig. 4, C and D) (*45*). Antivaccine content producers could therefore increase their audience's exposure to their content by manipulating these reactions. For example, antivaccine advocates could promote content that might be otherwise difficult to find by reacting to it, with content producers selecting posts that are more likely to spur more influential emotional reactions, such as "sad" (*45*).

Starting in September 2020, Facebook reduced the weight of angry reactions in the newsfeed ranking algorithm to zero (*45*). Following this change, we found that the odds that a post in an antivaccine page elicited an angry reaction decreased to 49% (95% CI, 0.47 to 0.52) of prepolicy trends and to 10% (95% CI, 0.02 to 0.64) of provaccine levels. The odds of an angry reaction also decreased in antivaccine groups but increased in provaccine groups to 54% (95% CI, 0.49 to 0.59) and 177% (95% CI, 1.49 to 2.11) of prepolicy trends (fig. S5), respectively. This led to an overall relative decrease in antivaccine groups relative to provaccine groups (OR = 0.30; 95% CI, 0.25 to 0.37).

In antivaccine groups, the odds of all other reactions increased to 130% (95% CI, 1.24 to 1.37) of prepolicy trends—1.18 (95% CI, 1.12–1.24) times more than in provaccine groups (Table 1)—whereas, in pages, it increased to 143% (95% CI, 1.39 to 1.48) of prepolicy trends, although this increase was not significant when compared to provaccine pages (OR = 2.71; 95% CI, 0.20 to 36.62). Thus, it appears that antivaccine Facebook users found ways to circumvent the intent of changes to Facebook's algorithms, using them to promote exposure, and obtain access to antivaccine content. Together, these observations suggest that flexibility enabled by the architecture of social media platforms is an important consideration when attempting to mitigate harmful online behavior.

DISCUSSION

Our findings suggest that Facebook's policies may have reduced the number of posts in antivaccine venues but did not induce a sustained reduction in engagement with antivaccine content. Misinformation proportions both on and off the platform appear to have increased. Furthermore, it appears that antivaccine page administrators especially focused on promoting content that outpaced updates to Facebook's moderation policies: The largest increases appear to have been associated with topics falsely attributing severe vaccine adverse events and deaths to COVID-19 vaccines. Since engagement levels with antivaccine page content did not differ significantly from prepolicy trends, this potentially reflects vaccine-hesitant users' desire for more information regarding a novel vaccine at a time when specific false claims had not yet been explicitly debunked. This underscores a need to account for, and address, the forces driving users' engagement with—i.e., demand for—misinformative content.

Limitations to this study include that only public Facebook pages and groups were studied. We do not make claims about behaviors in private spaces. However, the data available through transparency tools such as CrowdTangle constitute a critical window into the largest, and most public, venues on the world's largest social media platform. Substantial prior work has shown that important misinformation about vaccines is often found in public data (*56*, *57*) in part because antivaccine advocates seek to recruit the vaccine hesitant in public forums. We cannot rule out the possibility that provaccine venues might have contained some sensationalist

stories of vaccine harm; however, these stories do not appear to have been sufficiently prevalent to have been detected by our topic models. Last, our data cannot distinguish between posts or engagements made by unique individuals or the same individuals repeatedly posting. However, this concern is mitigated by the fact that, in pages, only administrators may post, whereas users in any venue may only react to a given post once. Taken as a whole, our findings emphasize the critical need for platforms to continue to provide researchers with access to these public data.

Social media platforms are engineered systems (*58*) with explicit design goals. In Facebook's case, these goals include facilitating the formation of online communities (*59*), which appear to use Facebook's architecture to expand access to antivaccine content, undermining attempts to curtail misinformation. The flexibility afforded by Facebook's system architecture, in conjunction with its position in a broader media ecosystem, may therefore have facilitated unintended consequences, including increased exposure to misinformation and political polarization, with potential implications for offline behaviors.

Our results suggest that attempts to address misinformation must go beyond hard and soft content remedies to account for the flexibility afforded by system architecture. Just as the products of building architecture must conform to building codes to protect public health and safety, social media platform designers must consider the public health consequences of their architectural choices and perhaps even develop codes that are consistent with best scientific evidence for reducing online harms.

MATERIALS AND METHODS

Throughout this study, we analyzed Facebook pages and groups separately because they serve different functions: Pages are designed for marketing and brand promotion, and only page administrators may post in them. In contrast, any member may post in groups, which serve as a forum for members to build community and discuss shared interests. This study was deemed exempt from Institutional Review Board (IRB) review by the George Washington University Office of Human Research (IRB #180804).

Experimental design

The objective of this study was to answer the following three research questions:

1) Were Facebook's policies associated with a substantial decrease in public antivaccine content and engagement with remaining content?

2) Did misinformation decrease when these policies were implemented?

3) How might Facebook's system architecture have enabled or undermined these policies?

To answer these questions, we used a CITS design to compare the weekly attributes of Facebook posts in public antivaccine and provaccine pages and groups to prepolicy trends. We also compared antivaccine to provaccine pages and groups. Statistically significant differences indicate that it is more likely to have been Facebook's policies and not some unrelated events, which caused the observed changes.

Research question 1: Did antivaccine content and engagement decrease?

To answer research question 1, we measured the following:

A) Weekly counts of the total number of posts in antivaccine and provaccine pages and groups.

B) User engagement with these posts.

Calculating engagement

We calculated engagement with each post as the sum of all likes, shares, comments, and other reactions (love, wow, haha, sad, angry, and care) reported by CrowdTangle. Facebook's algorithms reportedly use a weighted sum of these engagements when prioritizing content within users' newsfeeds (*44*).

Research question 2: Did misinformation decrease?

To answer research question 2, we examined outcome variables reflecting potential adverse consequences of Facebook's policies.

A) On the basis of the hypothesis that policies may have targeted known false claims without adapting to novel misinformation specific to COVID-19 vaccines (*22*), we examined the weekly fraction of posts discussing topics that were potentially prohibited under Facebook's policies.

B) On the basis of the hypothesis that posts on Facebook may have increasingly included links to misinformation on other platforms (*23*, *49*), we examined the daily fraction of web links pointing to known misinformative, but also politically polarized, websites.

Research question 3: Did Facebook's architecture help or hinder removal efforts?

Flexible systems enable users to access interdicted content via alternate pathways, even if some paths were removed (*34*–*36*). We observe that Facebook's architecture is a layered hierarchy made up of pages, groups, and users (Fig. 4) (*34*, *36*). This architecture may afford flexibility by providing several types of alternate paths. To answer research question 3, we examined outcome variables reflecting each layer in this hierarchy.

A) In the top layer, page administrators can collaborate to share followers and content by linking to one another's pages, enabling users to discover aligned pages and content. On the basis of the hypothesis that Facebook's policies might have eroded this capacity, we examined the weekly fraction of links pointing between pages in our dataset with the same stance (e.g., between antivaccine pages).

B) In the middle layer, groups and pages can engage in prohibited coordinated inauthentic behavior (*54*) by sharing the same content simultaneously. Thus, if one post is removed, then copies of it will still be present on the platform. We therefore examined whether antivaccine pages and groups routinely posted the same content near simultaneously—i.e., within 33 s of each other (see the "Middle layer: Coordinated link-sharing" section)—and whether the proportion of these near-simultaneous link pairs decreased after Facebook's policy.

C) In the bottom layer, Facebook's algorithms promote content in users' newsfeeds that has generated meaningful social interaction (*45*) as measured, in part, by a weighted sum of reactions (e.g., like, love, angry, etc.) that this content has spurred (Fig. 4, C and D) (*45*). Starting in September 2020, Facebook reduced the weight of angry reactions in the newsfeed ranking algorithm to zero because they were associated with toxic content and misinformation (*45*). We

therefore examined weekly proportions of angry and other reactions in provaccine and antivaccine pages and groups, based on the hypothesis that antivaccine content producers might have increased their audience's exposure to corresponding content by substituting the angry reaction for others.

Data

We downloaded data from CrowdTangle, a public insights tool owned and operated by Facebook. All application programming interfaces (APIs) that return data to researchers—including those that generate data from social media platforms such as Facebook, Twitter, and Reddit—are, to some extent, "black boxes" because they use proprietary algorithms when returning data (*60*). We opted to use CrowdTangle data because CrowdTangle was explicitly designed to mitigate this concern: It has been called the most effective transparency tool in the history of social media (*38*, *39*, *61*). In addition, we conducted several robustness checks to further mitigate the influence that any potential changes to Facebook or CrowdTangle's APIs might have had on our conclusions. Last, we note that CrowdTangle tracks interactions with public content from Facebook pages and groups. Similar to all social media platform APIs, CrowdTangle does not include activity on private accounts or posts made visible only to specific groups of followers; therefore, we do not make claims about activity within these private spaces.

Procedure for identifying vaccine-related pages and groups

Similar to prior work (*62*, *63*) that relies upon iterative procedures to reduce biases associated with keyword selection:

1) We first identified a large set of pages and groups that mentioned vaccines at least once within the most recent ~300,000 posts. We did not include venues with earlier timestamps since our aim was to capture activity in the most prolific Facebook groups in the lead-up to Facebook's policy implementation. To do so, we searched CrowdTangle on 15 November 2020, identifying and downloading all posts containing at least one keyword from the following list: vaccine, vaxx, vaccines, vaccination, vaccinated, vax, vaxxed, vaccinations, and jab. Several of these posts contained content pertaining to guns and to pet and other animal vaccines. Thus, we ran a second search excluding posts containing the following keywords: gun, dog, cat, foster, adopt, shelter, vet, kennel, pet, chicken, livestock, kitten, puppy, paw, and cow. We conducted this search on 15 November 2020 and retrieved the 299,981 most recent page posts and 299,904 group posts meeting search criteria before hitting CrowdTangle's download limit, with the earliest posts timestamped 7 September 2020 for pages and 1 July 2020 for groups. This procedure yielded 73,438 pages and 57,485 groups. To ensure that we only retained venues that routinely discussed vaccines, we retained all pages and groups whose name contained at least one of the strings "vacc," "vax," or "jab," or which posted very frequently about vaccines: i.e., in the top percentile—at least 44 times for pages and 58 times for groups. This procedure yielded 1231 pages and 773 groups.

2) Several of the venues generated in step 1 were news organizations that did not primarily focus on vaccination. We therefore further narrowed down our initial list as follows: We retrieved as many posts as possible from these venues yielding the 299,994 most recent page posts and 299,969 group posts before hitting CrowdTangle's download limit. As above, we did not collect earlier posts since our aim was to identify venues that actively posted about vaccination in the lead-up to the policy's implementation. The earliest post from these venues was timestamped 8 November 2020 for pages and 13 November 2020 for groups. We retained all venues for which at least 20% of posts retrieved contained at least one word containing vacc or vax. We selected this 20% threshold by inspection, and our results were insensitive to changing it (relaxing the threshold yielded more groups and pages that were characterized as "other" in step 3, which were not included in our analysis. All groups and pages were checked for relevance by two authors—A.M.J. and J.G.).

3) On 15 November 2020, we retrieved all posts from the venues identified in step 2 for a 12-month period starting on 15 November 2019, forward. We did not exceed the download limit for these venues. We repeatedly collected content from these venues from 30 November 2020 to 28 February 2022 (see table S4). Qualitatively, we found that these posts focused on general vaccine content, adhering closely to existing typologies of provaccine and antivaccine topics (*64*, *65*). We therefore manually annotated these venues as provaccine, antivaccine, or other. Two independent annotators (J.G. and A.M.J.) manually assessed each group and page following a two-tiered coding scheme (*65*), achieving high reliability (Cohen's $\kappa = 0.88$; 95% CI, 0.85 to 0.92). Categorization was based on content shared in the "about" section of each venue. When this section was left blank, annotators considered the venue's title, any imagery used, and recent posts to decide. All venues were double-coded, with disagreements discursively reconciled. We retained all venues that were labeled as provaccine or antivaccine.

This procedure is summarized in fig. S6.

Statistical analysis

To answer question 1 (A), we extracted a total of eight weekly time series. The first four of these time series reflect the total number of vaccine-related posts in (i) antivaccine pages, (ii) antivaccine groups, (iii) provaccine pages, and (iv) provaccine groups. To answer question 1 (B), the next four time series examined engagements with these posts, since Facebook's algorithms uses a sum of different engagement types (likes, shares, etc.) when prioritizing content within users' newsfeeds, meaning that posts receiving many engagements are more likely to be seen (*44*). We applied a logarithm transform to all counts to correct for data skew.

The prepolicy period was defined as 15 November 2019 to 15 November 2020—the same week that Facebook removed Stop Mandatory Vaccination—a large page with over 360,000 members (*40*) and 2 weeks before Facebook publicly committed to removing misinformation about COVID-19 vaccines (*41*). The postpolicy period was defined as the following week to 28 February 2022. We conducted robustness checks to examine the effects of shifting the policy date as early as 19 August 2020 (when Facebook committed to removing calls to violence related to the QAnon conspiracy theory) and as late as 8 February 2021 (when Facebook stated that they would remove all vaccine-related misinformation).

CITS analysis

We examined changes to the weekly number of posts, as well as engagements with those posts, using a CITS design with a nonequivalent control group—one of the strongest quasi-experimental designs available (*66*). This design enabled us to estimate the effects of changes to Facebook's policies. Specifically, any changes to observed data affecting antivaccine content that are not due to

Facebook's policies—e.g., external news about vaccine trials—are expected to affect both provaccine and antivaccine groups and pages since they are both focused on vaccination. We compared the year before 15 November 2020 to the remainder of the dataset, enabling us to estimate the effects of Facebook's hard content remedies on antivaccine content.

We conducted interrupted time-series analyses using autoregressive integrated moving average (ARIMA) models fit to weekly sums of posts and engagements from 17 November 2019 to 15 November 2020. To control for the formation of new pages and groups between 15 November 2019 and 15 November 2020, we divided weekly post and engagement counts by the total number of weekly venues in each dataset before fitting ARIMA models. We fit all ARIMA models to prepolicy data using the auto_arima function in the pmdarima Python package (*67*). When time series were not stationary, as determined by an augmented Dickey-Fuller test, data were detrended using differencing. We selected the number of autoregressive and moving average terms employing a parallel grid search and otherwise using pmdarima's default settings.

We used these models to generate counterfactual projections for weekly posts or engagements that would have been present assuming no hard remedies. We calculated the percent difference between these counterfactual projections and observed post and engagement counts. We consider a policy to have been effective if it consistently reduced content beyond the 95% confidence bounds of these projections.

Provaccine venues make an ideal nonequivalent control group because, similar to antivaccine venues, they contain users who are motivated to post about vaccines and would therefore respond to exogenous factors, such as the news cycle, in the same way; however, platforms' policies were not designed to target provaccine content. A statistically significant difference in antivaccine, but not provaccine, venues, or between antivaccine and provaccine venue effect sizes, indicates that it is more likely to have been Facebook's policies, and not some contemporaneous event, which caused the observed change.

According to the CITS approach, Facebook's content and account removal policies in antivaccine pages and groups were effective if they caused proportionally greater deviations from prepolicy trends than in provaccine pages and groups. We calculated prepolicy trends by fitting ARIMA (*68*) models to weekly retrospective data covering the prepolicy period, and collected on 15 November 2020. This model is specified as

$$\Theta(L)^p \Delta^d \Delta_S^D y_t = \Phi(L)^q \Delta^d \Delta_S^D \varepsilon_t \quad (1)$$

where y_t is the quantity being predicted (e.g., the daily number of posts), ε_t is the error at time t, $\Theta(L)^p$ is a p-order polynomial function of L capturing autoregressive terms (i.e., $L^n y_t = y_{t-n}$), and $\Phi(L)^q$ is a q-order polynomial function of L capturing moving average terms (i.e., $L^n \varepsilon_t = \varepsilon_{t-n}$). In addition, $y_t^{[d]} = \Delta^d y_t = y_t^{[d-1]} - y_{t-1}^{[d-1]}$, where $y_t^{[0]} = y_t$ is the order of differencing.

To ensure stationarity and trend stationarity of our input data, we applied augmented Dickey-Fuller tests to each input dataset and applied differencing if these tests indicated that it was necessary. We next determined the order of autoregressive (AR) and moving average (MA) terms using the auto_arima function as implemented in the pmdarima Python package, selecting the model that minimized the Akaike information criterion after performing an exhaustive grid search using default settings.

Having fit ARIMA models, we used the procedure recommended by Fanshawe *et al.* (*69*) to compare postpolicy data to prepolicy trends by calculating 1000 forecasted time series using the "simulate" function in the statsmodels Python package. For each forecasted time series, we calculated $M_k = \sum_{t=n+1}^{n+k} \hat{y}_t / k$, where n is the total number of prepolicy days and k is the total number of postpolicy days. Thus, M_k represents the average of the forecasted values over the postpolicy period. Given 1000 such forecasts, we were able to calculate the mean, $\hat{m}_k$, and SD, $\hat{s}_k$, of M_k. Given a vector of postpolicy data, y_t, we calculated a Z statistic for each postpolicy time series as $Z = \frac{\sum_{t=n+1}^{n+k} y_t/k - \hat{m}_k}{\hat{s}_k}$, where the numerator of this quantity reflects the average difference between actual and simulated postpolicy means and the denominator reflects the SD of this difference. Consequently, the 95% CIs for this quantity can be estimated as

$$CI = \sum_{t=n+1}^{n+k} y_t/k - \hat{m}_k \pm \mathcal{N}\left(1+\frac{\alpha}{2}\right) * \hat{s}_k \quad (2)$$

where $\mathcal{N}\left(1-\frac{\alpha}{2}\right)$ is the inverse cumulative distribution function for a type I error rate of α. For this study, α = 0.05, and $\mathcal{N}\left(1-\frac{\alpha}{2}\right) = 1.96$ unless otherwise specified (e.g., when correcting for multiple comparisons).

We calculated the effect of Facebook's policies on antivaccine pages and groups relative to provaccine pages and groups as

$$Z = \frac{\left(\sum_{t=n+1}^{n+k} y_{t,\text{anti}}/k - \hat{m}_{k,\text{anti}}\right) - \left(\sum_{t=n+1}^{n+k} y_{t,\text{pro}}/k - \hat{m}_{k,\text{pro}}\right)}{\sqrt{(\hat{s}_{k,\text{anti}})^2 + (\hat{s}_{k,\text{pro}})^2}} \quad (3)$$

with the numerator of this quantity reflecting the average relative difference and the denominator reflecting the SD of this difference. We conducted analyses on post and engagement count data using log transformations; thus, exponentiating these average differences yields a risk ratio. Similarly, analyses of proportions were conducted using logit transformations to account for multinomial variance-covariance structures for which exponentiating average daily differences yields ORs. Last, analyses conducted on Likert-scale data were not transformed.

A statistically significant difference between antivaccine and provaccine effect sizes indicates that it is more likely to have been Facebook's policies, and not some contemporaneous event, which caused the observed change. In contrast, statistically significant differences in both antivaccine and provaccine posts may indicate either the effects of some exogenous event or, alternatively, that Facebook's policies affected both types of content. By using this approach, we were able to make inferences about the effects of Facebook's policies that would not be possible using antivaccine data alone.

Unlike approaches such as the difference-in-differences design (*70*), this approach does not require us to assume that the relationship between Facebook's policies and activity in provaccine and antivaccine pages and groups is constant over time or to assume a

"burn-in" period of arbitrary length during which Facebook's policies were being rolled out. Rather, this approach allows us to assess the net effect of Facebook's policies throughout the postpolicy period. This is especially important because Facebook's own press releases indicated that users should expect a delay between Facebook's policy announcements and a fully realized content removal policy. Furthermore, this model does not require us to assume that Facebook implemented their policies consistently throughout the postpolicy period but may have instead engaged in multiple waves of content and account removal in responses to external events (e.g., media attention).

Statistical analyses were performed using Python's pmdarima (*71*) and statsmodels (*72*) packages. All tests were two-sided and a P value of 0.05 or less was considered statistically significant.

Structural topic model analysis

To answer question 2 (A), we extracted the text from each Facebook post by combining the "Message," "Image Text," "Link Text," and "Description" fields returned by CrowdTangle. We first identified unique English posts using the langdetect Python package (*73*), yielding 268,875 antivaccine posts and 76,954 provaccine posts collected in the pages and groups identified 15 November 2020. We next removed non-English posts and converted all English text to lowercase. Next, we removed all numerals, punctuation, symbols, and web links.

Structural topic models (STMs) were fit using the R Project for Statistical Computing's stm (*74*) software package. Data were used as input to the STM algorithm (*74*), using the Mimno and Lee algorithm (*75*) to automatically select the number of topics. STMs enable a comparison between antivaccine and provaccine aspects of the same topic, allowing us to use provaccine topics as a nonequivalent control group. We used the resulting model to extract several indicators of topic interpretation, including word clouds depicting the most frequent words for each topic and example posts with the highest fraction of each topic (see tables S1 and S2). We assigned descriptive labels to each topic based on these indicators. Next, we used these labels to assign each topic to descriptive categories following a widely used and validated typology of vaccine-related online content (*56*, *64*, *65*). In addition, we determined whether these descriptive labels matched content that Facebook's COVID-19 and Vaccine Policy listed as misinformative or otherwise prohibited (*76*). Specifically, three coders (A.M.J., J.G., and D.A.B.) independently assigned thematic labels to the 61 topics based off of the word clouds generated by the model (table S1). We next assessed similarity between labels, indicating matches that were very close, similar, and not at all close. For example, J.G.'s "Alternative Medicine," A.M.J.'s "Natural Remedies," and D.A.B.'s "Homeopathy" were considered very close labels because they captured the same underlying theme. Meanwhile, while J.G. and D.A.B. agreed on "Vaccine Mandates," A.M.J.'s label of "Gov't Overreach" was only considered similar because, while they captured similar underlying themes, they focused on different dimensions. Of the 61 topics, coders' labels were considered very close, or similar matches for 53 (87%) topics. Overall, there was very high agreement among all three raters, Fleiss' $\kappa = 0.895$, $P < 0.001$. Among the topics where raters did not agree, there were five topics where two of three coders agreed, but the third coder did not align. There were three topics where there was no clear agreement between any of the coders. For these eight topics, we engaged in a second round of coding, based on the full text of the top 10 provaccine posts and top 10 antivaccine posts for each topic. With this additional data, we revisited our codes and discursively reached agreement. After reaching consensus, the final labels were applied to each topic.

Next, the same three coders each independently assessed whether content from each of the 61 topics were prohibited under Facebook's content policies. Options included yes for potentially prohibited and no for likely not prohibited. For topics labeled as prohibited, we also included justifications for which policies were likely to be violated. Of the 61 topics, coders agreed on 52 topics. Overall, there was substantial agreement among all three raters, Fleiss' $\kappa = 0.735$, $P < 0.001$. For the nine topics on which raters did not agree, the team discussed responses, discursively reached agreement, and applied a final decision.

After labeling each topic, we calculated proportions of each topic in each English post, applying a logistic transform to control for floor and ceiling effects. We extracted time series by averaging weekly proportions of each topic in antivaccine and provaccine pages and groups, respectively. Next, we fit ARIMA models to pre-policy trends. We corrected for potential type I error due to multiple comparisons using the Benjamini-Hochberg procedure (table S3) (*77*).

Misinformative and polarized URL analysis

Facebook posts frequently contain links to external websites. To answer question 2 (B), we extracted these links and determined whether they pointed to misinformative and polarized sources by comparing them to validated, publicly available lists of these sources (*78*). Specifically, for each dataset, we extracted all URLs in the "Link" field returned by CrowdTangle or the "Final Link" field if it was nonempty (meaning that the "Link" field used a URL shortener). Next, using the TLDExtract Python module (*79*), we extracted the top-level domain (TLD) and suffix for each URL (for example, the TLD of www.example.com/this-is-an-example.html is example.com).

A large body of prior work (*78*, *80*–*83*) shows that posts with links to noncredible sources may serve as a proxy for misinformation (*46*, *84*, *85*). We next calculated the weekly proportion in provaccine and antivaccine venues of all posts with a TLD listed on iffy.news (*86*)—a list of publishers identified by MediaBiasFactCheck.com as having "low" or "very low" factual reporting scores—on 30 April 2022. MediaBiasFactCheck ratings are known to be strongly correlated with several other URL credibility ratings (*78*, *81*). We also calculated the weekly proportion of engagements with low-credibility URLs by weighting each post containing an "Iffy" off-platform URL by the total number of engagements with it. We conducted interrupted time-series ARIMA analyses on these weekly averages after applying logit transforms to correct for floor and ceiling effects. We did not analyze provaccine content because Iffy URLs were largely absent from provaccine venues.

MediaBiasFactCheck.com also scores websites by their partisan bias ranging from far right (−4) to far left (*4*). We collated a list of 1314 publishers that had been scored by MediaBiasFactCheck.com as of 3 February 2022. We next calculated the weekly proportion in provaccine and antivaccine venues of all posts with a TLD identified by MediaBiasFactCheck.com as having "right," "far right," "extreme right," "left," "far left," or "extreme left" bias scores. We also calculated the weekly proportion of engagements with polarized URLs by

weighting each post containing a polarized off-platform URL by the total number of engagements with it. We conducted interrupted time-series ARIMA analyses on these weekly averages after applying logit transforms to correct for floor and ceiling effects. We did not analyze provaccine content because polarized URLs were absent from provaccine venues during several weeks.

Last, we calculated the weekly average "Bias Rating" for all rated links and examined how this average changed over time in both provaccine and antivaccine venues. Links to these publishers made up 34 and 17% of off-platforms links to pages and groups, among venues identified 15 November 2020, garnering 41 and 19% of engagements, respectively. Similarly, links to these publishers made up 34 and 15% of off-platforms links to pages and groups, among venues identified 21 July 2021, garnering 55 and 14% of engagements.

Top layer: Building Facebook page networks

To answer research question 3 (A), we examined whether Facebook's new policies reduced links between the antivaccine and provaccine pages in our sample. (The number of links to, and between, groups was zero in most weeks.) We identified all posts containing a link starting with www.facebook.com. We then calculated the proportion of all posts containing a link that pointed to antivaccine and provaccine pages in each of our samples. Specifically, we identified and extracted the unique numerical Facebook ID and username for each provaccine and antivaccine page in each dataset. We considered a source page to be linked to a target page if a URL posted on the source page began with www.facebook.com/<Facebook ID>/ or www.facebook.com/<account name>/. We next calculated the weekly proportion of all posts that pointed from antivaccine pages to other antivaccine pages (excluding self-links) and from provaccine pages to other provaccine pages. We next conducted CITS analyses on these weekly proportions. After applying a logit transform to our data to control for floor and ceiling effects, we fit ARIMA models to prepolicy data, comparing postpolicy data to model projections using the same techniques and transforms applied to answer research question 2 (B). Since pages largely overlapped between samples identified 15 November 2020 and 21 July 2021, we combined these data since doing so allowed us to also examine links between pages that were in separate datasets (separate analyses of these datasets may be found in fig. S1). We also used the links that we extracted to construct unweighted networks. In practice, these networks qualitatively resemble prior work based on mutual "likes" between pages but which currently require access to Facebook's commercial APIs to replicate at scale (*53*, *87*). Similar to this prior work, our networks were displayed using a force-directed layout algorithm (*88*).

Middle layer: Coordinated link-sharing

To answer research question 3 (B), we observe that Facebook's community standards disallow "us[ing] fake accounts, artificially boost[ing] the popularity of content, or engag[ing] in behaviors designed to enable other violations under our Community Standards" (*54*). Prior works (*89*, *90*) suggest that this type of coordinated inauthentic behavior may be detected under the assumption that "near-simultaneous link sharing" is a signal of coordination (*90*–*93*). Building upon prior work (*90*), we operationalized near-simultaneous link sharing in a manner that was intended to be robust to the specific query being used. We conducted three "blank search" queries between 30 March and 31 March 2020 on CrowdTangle to identify the ~300,000 most recent posts each for pages and groups available on the platform, combining across pages and groups. We calculated the time difference in seconds between each successive share of the same URLs. To distinguish between coordinated and uncoordinated behaviors, we modeled the distribution of these interarrival times as a mixture of exponential distributions with components corresponding to "near simultaneous" sharing and "nonsimultaneous" sharing. We used the expmixture-model package (*94*) in Python to fit these exponential mixture models to interarrival time data derived from three blank searches conducted on CrowdTangle between 30 March and 31 March 2021. We found that several model goodness-of-fit measures converged on a two-component distribution for each dataset (see table S5).

We next examined the goodness of fit of these estimated distributions by comparing them each one to its respective dataset (see the Supplementary Materials) and using Kolmogorov-Smirnov tests. Although each test detected a significant difference (d_1 = 0.12, d_2 = 0.13, and d_3 = 0.13, $P < 0.001$ in all cases), our large dataset sizes (n_1 = 87,000, n_2 = 85,346, and n_3 = 93,075) mean that we are powered to detect very small differences. Furthermore, our model underestimates the likelihood of near-simultaneous sharing for very small time differences (between 0 and 15 s), meaning that our estimated threshold is conservative (see fig. S7).

The best fitting model was made up of two components, with mean interarrival times of $\mu_{\text{near-simultaneous}}$ = 9.95 s and $\mu_{\text{nonsimultaneous}}$ = 227.45 s, meaning that a URL shared by two venues in under 33 s is more likely to have originated from the near-simultaneous component than the nonsimultaneous component. This number is comparable to thresholds defined heuristically in prior work (*90*–*93*). We considered venues to be routinely coordinated if their empirical frequency of near simultaneous links significantly exceeded 13.34%—the expected likelihood that links were drawn from the nonsimultaneous distribution—using binomial tests. Venues were linked if they were significantly coordinated at the $P < 0.05$ level after controlling for multiple comparisons using the Holm-Bonferroni procedure. We also tested several threshold values ranging from 25 to 41 s and found that results qualitatively replicated. We also calculated the proportion of all link pairs that were classified as near-simultaneous and, after applying a logit transform, fit ARIMA models to prepolicy data, comparing postpolicy data to prepolicy projections (see fig. S3) using the same techniques and transforms applied to answer research question 2 (B). We were unable to calculate ARIMA models for provaccine data since coordinated links were absent for several months during the spring and summer of 2020. However, starting in Fall 2020, we observed an increase in provaccine coordinated behavior.

Bottom layer: Angry and other reactions

To answer research question 3 (C), for each dataset, we calculated the weekly proportion of engagements that were reportedly (*45*) given zero weight (angry) and those that were positively weighted by Facebook's newsfeed algorithm (likes, love, sad, haha, wow, and care) and conducted interrupted time-series analyses on these proportions. We fit ARIMA models to logit-transformed prepolicy data, comparing postpolicy data to prepolicy projections (see fig. S5) using the same techniques and transforms applied to answer research question 2 (B).

Robustness checks

We conducted several robustness checks to evaluate whether our results were sensitive to modeling assumptions. Specifically, we examined the sensitivity of our results to the following:

1) Selection of policy date: We replicated results of our CITS analysis using a range of policy dates extending from 17 August 2020 to 8 February 2021. Results of this analysis suggest that these alternative dates do not provide better explanations of the data than our reference date of 18 November 2020.

2) Engagement aggregation method: We replicated results of our engagement count analysis, examining types of engagements (comments, shares, and reactions) separately. Results of this analysis suggest that our primary results largely replicated across engagement types and especially for reactions.

3) Topic modeling technique: We replicated results of our STM analysis using other topic modeling techniques. Results of this analysis suggest that our findings are not sensitive to the specific topic model used.

4) Data collection date: We identified a second set of anti- and provaccine pages and groups on 28 July 2021. We found that our results largely replicated the findings of the first set of pages and groups.

5) Data aggregation window: We examined whether our results changed when aggregating data by day rather than by week. We found that our results largely replicated across data aggregation windows.

Detailed results for all of these robustness checks are found in the Supplementary Materials.

Supplementary Materials

This PDF file includes:

Supplementary Text
Figs. S1 to S23
Tables S1 to S10
References

REFERENCES AND NOTES

1. J. D. West, C. T. Bergstrom, Misinformation in and about science. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e1912444117 (2021).
2. W.-Y. S. Chou, A. Oh, W. M. P. Klein, Addressing health-related misinformation on social media. *JAMA* **320**, 2417–2418 (2018).
3. 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).
4. B. Silverman, R. Mac, J. Lytvynenko, "How Facebook failed to prevent stop the steal," *BuzzFeed News*, 22 April 2021; www.buzzfeednews.com/article/craigsilverman/facebook-failed-stop-the-steal-insurrection.
5. A. Gowen, "As mob lynchings fueled by WhatsApp messages sweep India, authorities struggle to combat fake news," *Washington Post*, 2 July 2018; www.washingtonpost.com/world/asia_pacific/as-mob-lynchings-fueled-by-whatsapp-sweep-india-authorities-struggle-to-combat-fake-news/2018/07/02/683a1578-7bba-11e8-ac4e-421ef7165923_story.html.
6. P. Ball, A. Maxmen, The epic battle against coronavirus misinformation and conspiracy theories. *Nature* **581**, 371–374 (2020).
7. A. Tyson, C. Johnson, C. Funk, "U.S. public now divided over whether to get COVID-19 vaccine," *Pew Research Center Science & Society*, 17 September 2020; www.pewresearch.org/science/2020/09/17/u-s-public-now-divided-over-whether-to-get-covid-19-vaccine/.
8. "Senators Klobuchar, Baldwin, Peters urge tech industry leaders to combat coronavirus vaccine misinformation," *U.S. Senator Amy Klobuchar*, 25 January 2021; www.klobuchar.senate.gov/public/index.cfm/2021/1/senators-klobuchar-baldwin-peters-urge-tech-industry-leaders-to-combat-coronavirus-vaccine-misinformation.
9. "Call to action: CSIS-LSHTM high-level panel on vaccine confidence and misinformation," 19 October 2020; www.csis.org/analysis/call-action-csis-lshtm-high-level-panel-vaccine-confidence-and-misinformation.
10. O. Papakyriakopoulos, E. Goodman, The impact of Twitter labels on misinformation spread and user engagement: Lessons from Trump's election tweets, in *Proceedings of the ACM Web Conference 2022* (Association for Computing Machinery, 2022), pp. 2541–2551; https://dl.acm.org/doi/10.1145/3485447.3512126.
11. J. Nassetta, K. Gross, "State media warning labels can counteract the effects of foreign disinformation," *Harvard Kennedy School Misinformation Review*, 30 October 2020.
12. J. Gu, A. Dor, K. Li, D. A. Broniatowski, M. Hatheway, L. Fritz, L. C. Abroms, The impact of Facebook's vaccine misinformation policy on user endorsements of vaccine content: An interrupted time series analysis. *Vaccine* **40**, 2209–2214 (2022).
13. G. Pennycook, Z. Epstein, M. Mosleh, A. A. Arechar, D. Eckles, D. G. Rand, Shifting attention to accuracy can reduce misinformation online. *Nature* **592**, 590–595 (2021).
14. J. Roozenbeek, A. L. J. Freeman, S. van der Linden, How accurate are accuracy-nudge interventions? A preregistered direct replication of Pennycook et al. (2020). *Psychol. Sci.* **32**, 1169–1178 (2021).
15. E. Goldman, Content moderation remedies. *Mich. Tech. L. Rev.* **28**, 1 (2021).
16. S. Jhaver, C. Boylston, D. Yang, A. Bruckman, Evaluating the effectiveness of deplatforming as a moderation strategy on Twitter. *Proc. ACM Hum. Comput. Interact.* **5**, 1–30 (2021).
17. S. Ali, M. H. Saeed, E. Aldreabi, J. Blackburn, E. De Cristofaro, S. Zannettou, G. Stringhini, Understanding the effect of deplatforming on social networks, in *13th ACM Web Science Conference 2021* (Association for Computing Machinery, 2021), pp. 187–195; https://doi.org/10.1145/3447535.3462637.
18. E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, E. Gilbert, You can't stay here. *Proc. ACM Hum.-Comput. Interact.* **1**, 1–22 (2017).
19. J. B. Bak-Coleman, I. Kennedy, M. Wack, A. Beers, J. S. Schafer, E. S. Spiro, K. Starbird, J. D. West, Combining interventions to reduce the spread of viral misinformation. *Nat. Hum. Behav.* **6**, 1372–1380 (2022).
20. D. A. Broniatowski, M. Dredze, J. W. Ayers, "First do no harm": Effective communication about COVID-19 vaccines. *Am. J. Public Health* **111**, 1055–1057 (2021).
21. R. Gorwa, R. Binns, C. Katzenbach, Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data Soc.* **7**, 10.1177/2053951719897945, (2020).
22. T. Gillespie, Content moderation, AI, and the question of scale. *Big Data Soc.* **7**, 10.1177/205395172094, (2020).
23. M. H. Ribeiro, S. Jhaver, S. Zannettou, J. Blackburn, E. De Cristofaro, G. Stringhini, R. West, Do platform migrations compromise content moderation? Evidence from r/The_Donald and r/Incels. *Proc. ACM Hum. Comput. Interact.* **5**, 1–24 (2021).
24. D. R. Thomas, L. A. Wahedi, Disrupting hate: The effect of deplatforming hate organizations on their online audience. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2214080120 (2023).
25. "Senate Bill 7072 (2021) - The Florida Senate"; www.flsenate.gov/Session/Bill/2021/7072/.
26. "87(2) HB 20 - Senate committee report version - Bill text"; https://capitol.texas.gov/tlodocs/872/billtext/html/HB00020S.htm.
27. D. A. Broniatowski, V. F. Reyna, To illuminate and motivate: A fuzzy-trace model of the spread of information online. *Comput. Math. Organ. Theory* **26**, 431–464 (2020).
28. S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online. *Science* **359**, 1146–1151 (2018).
29. Z. Epstein, N. Sirlin, A. Arechar, G. Pennycook, D. Rand, The social media context interferes with truth discernment. *Sci. Adv.* **9**, eabo6169 (2023).
30. V. F. Reyna, Social media: Why sharing interferes with telling true from false. *Sci. Adv.* **9**, eadg8333 (2023).
31. M. Cinelli, G. De Francisci Morales, A. Galeazzi, W. Quattrociocchi, M. Starnini, The echo chamber effect on social media. *Proc. Natl. Acad. Sci.* **118**, e2023301118 (2021).
32. M. D. Vicario, G. Vivaldo, A. Bessi, F. Zollo, A. Scala, G. Caldarelli, W. Quattrociocchi, Echo chambers: Emotional contagion and group polarization on Facebook. *Sci. Rep.* **6**, 37825 (2016).
33. R. E. Robertson, J. Green, D. J. Ruck, K. Ognyanova, C. Wilson, D. Lazer, Users choose to engage with more partisan news than they are exposed to on Google Search. *Nature* **618**, 342–348 (2023).
34. D. A. Broniatowski, J. Moses, Measuring flexibility, descriptive complexity, and rework potential in generic system architectures. *Syst. Eng.* **19**, 207–221 (2016).
35. J. Moses, Flexibility and its relation to complexity and architecture, in *Complex Systems Design & Management*, M. Aiguier, F. Bretaudeau, D. Krob, Eds. (Springer, 2010), pp. 197–206; http://link.springer.com/chapter/10.1007/978-3-642-15654-0_14.
36. D. A. Broniatowski, Flexibility due to abstraction and decomposition. *Syst. Eng.* **20**, 98–117 (2017).

37. "Facebook reports fourth quarter and full year 2020 results"; https://investor.fb.com/investor-news/press-release-details/2021/Facebook-Reports-Fourth-Quarter-and-Full-Year-2020-Results/default.aspx.
38. CrowdTangle Team, *CrowdTangle* (Menlo Park, CA, USA, 2021; List IDs: 1475046, 1475047, 1584315, 1584316).
39. B. Smith, "A former Facebook executive pushes to open social media's 'black boxes'," *The New York Times*, 2 January 2022; www.nytimes.com/2022/01/02/business/media/crowdtangle-facebook-brandon-silverman.html.
40. A. Sulleyman, "Facebook bans one of the anti-vaccine movement's biggest pages for violating QAnon rules," *Newsweek*, 18 November 2020; www.newsweek.com/facebook-bans-anti-vaccine-group-violating-qanon-rules-1548408.
41. K.-X. Jin, "Keeping people safe and informed about the coronavirus," *Meta*, 18 December 2020; https://about.fb.com/news/2020/12/coronavirus/.
42. K.-X. Jin, "Reaching billions of people with COVID-19 vaccine information," *Meta*, 8 February 2021; https://about.fb.com/news/2021/02/reaching-billions-of-people-with-covid-19-vaccine-information/.
43. "What role does a Page have in a group?" Facebook Help Center; www.facebook.com/help/2003297226584040.
44. W. Oremus, C. Alcantara, J. B. Merrill, A. Galocha, "How Facebook shapes your feed," *Washington Post*, 26 October 2021; www.washingtonpost.com/technology/interactive/2021/how-facebook-algorithm-works/.
45. J. B. Merrill, W. Oremus, "Five points for anger, one for a 'like': How Facebook's formula fostered rage and misinformation," *Washington Post*, 26 October 2021; www.washingtonpost.com/technology/2021/10/26/facebook-angry-emoji-algorithm/.
46. 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).
47. C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, F. Menczer, The spread of low-credibility content by social bots. *Nat. Commun.* **9**, 4787 (2018).
48. S. Schechner, J. Horwitz, E. Glazer, "How Facebook hobbled Mark Zuckerberg's bid to get America vaccinated," *Wall Street Journal*, 17 September 2021; www.wsj.com/articles/facebook-mark-zuckerberg-vaccinated-11631880296.
49. T. Mitts, N. Pisharody, J. Shapiro, Removal of anti-vaccine content impacts social media discourse, in *14th ACM Web Science Conference 2022* (Association for Computing Machinery, 2022), pp. 319–326; https://doi.org/10.1145/3501247.3531548.
50. Y. Zhou, M. Dredze, D. A. Broniatowski, W. D. Adler, Elites and foreign actors among the alt-right: The gab social media platform. *First Monday* **24**, 10.5210/fm.v24i9.10062, (2019).
51. M. Trujillo, M. Gruppi, C. Buntain, B. D. Horne, What is BitChute? Characterizing the "Free Speech" alternative to YouTube, in *Proceedings of the 31st ACM Conference on Hypertext and Social Media* (ACM, 2020), pp. 139–140; https://dl.acm.org/doi/10.1145/3372923.3404833.
52. K. Starbird, A. Arif, T. Wilson, Disinformation as collaborative work. *Proc. ACM Hum. Comput. Interact.* **3**, 1–26 (2019).
53. N. F. Johnson, N. Velásquez, N. J. Restrepo, R. Leahy, N. Gabriel, S. El Oud, M. Zheng, P. Manrique, S. Wuchty, Y. Lupu, The online competition between pro- and anti-vaccination views. *Nature* **582**, 230–233 (2020).
54. "Inauthentic behavior," Transparency Center; https://transparency.fb.com/policies/community-standards/inauthentic-behavior/.
55. K. Hagey, J. Horwitz, "Facebook tried to make its platform a healthier place. It got angrier instead," *Wall Street Journal*, 15 September 2021; www.wsj.com/articles/facebook-algorithm-change-zuckerberg-11631654215.
56. D. A. Broniatowski, A. M. Jamison, N. F. Johnson, N. Velasquez, R. Leahy, N. J. Restrepo, M. Dredze, S. C. Quinn, Facebook pages, the "Disneyland" measles outbreak, and promotion of vaccine refusal as a civil right, 2009–2019. *Am. J. Public Health* **110**, S312–S318 (2020).
57. A. M. Jamison, D. A. Broniatowski, M. Dredze, A. Sangraula, M. C. Smith, S. C. Quinn, Not just conspiracy theories: Vaccine opponents and proponents add to the COVID-19 'infodemic' on twitter. *Harv. Kennedy Sch. Misinfo. Rev.* **1**, 10.37016/mr-2020-38, (2020).
58. O. L. De Weck, D. Roos, C. L. Magee, *Engineering Systems: Meeting Human Needs in a Complex Technological World* (MIT Press, 2011).
59. "Company info," *Meta*; https://about.facebook.com/company-info/.
60. R. Tromble, Where have all the data gone? A critical reflection on academic digital research in the post-API age. *Society* **7**, 205630512198892 (2021).
61. "Platform transparency: Understanding the impact of social media," *U.S. Senate Committee on the Judiciary*, 4 May 2022; www.judiciary.senate.gov/committee-activity/hearings/platform-transparency-understanding-the-impact-of-social-media.
62. G. King, P. Lam, M. E. Roberts, Computer-assisted keyword and document set discovery from unstructured text. *Am. J. Polit. Sci.* **61**, 971–988 (2017).
63. M. Dredze, D. A. Broniatowski, M. C. Smith, K. M. Hilyard, Understanding vaccine refusal: Why we need social media now. *Am. J. Prev. Med.* **50**, 550–552 (2016).
64. A. Kata, A postmodern Pandora's box: Anti-vaccination misinformation on the internet. *Vaccine* **28**, 1709–1716 (2010).
65. A. Jamison, D. A. Broniatowski, M. C. Smith, K. S. Parikh, A. Malik, M. Dredze, S. C. Quinn, Adapting and extending a typology to identify vaccine misinformation on Twitter. *Am. J. Public Health* **110**, S331–S339 (2020).
66. W. R. Shadish, T. D. Cook, D. T. Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (Wadsworth Cengage learning, 2002); https://pdfs.semanticscholar.org/f141/aeffd3afcb0e76d5126bec9ee860336bee13.pdf.
67. T. G. Smith, *et al.*, "pmdarima: ARIMA estimators for Python" (2017); www.alkaline-ml.com/pmdarima.
68. C. Chatfield, *The Analysis of Time Series* (Chapman and Hall/CRC, ed. 0, 2003); www.taylorfrancis.com/books/9780203491683.
69. T. R. Fanshawe, P. J. Turner, M. M. Gillespie, G. N. Hayward, The comparative interrupted time series design for assessment of diagnostic impact: Methodological considerations and an example using point-of-care C-reactive protein testing. *Diagn. Progn. Res.* **6**, 3 (2022).
70. J. B. Dimick, A. M. Ryan, Methods for evaluating changes in health care policy. *JAMA* **312**, 2401–2402 (2014).
71. "pmdarima: ARIMA estimators for Python — pmdarima 1.8.5 documentation"; http://alkaline-ml.com/pmdarima/.
72. S. Seabold, J. Perktold, statsmodels: Econometric and statistical modeling with Python, in *Proceedings of the 9th Python in Science Conference* (Austin, TX, 2010), pp. 10-25080.
73. M. M. Danilak, "langdetect: Language detection library ported from Google's language-detection," *GitHub*; https://github.com/Mimino666/langdetect.
74. M. E. Roberts, B. M. Stewart, D. Tingley, Stm: An R package for structural topic models. *J. Stat. Softw.* **91**, 1–40 (2019).
75. D. Mimno, M. Lee, Low-dimensional embeddings for interpretable anchor-based topic inference, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2014), pp. 1319–1328.
76. "COVID-19 and vaccine policy updates & protections," *Facebook Help Center*; www.facebook.com/help/230764881494641.
77. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. B. Methodol.* **57**, 289–300 (1995).
78. D. A. Broniatowski, D. Kerchner, F. Farooq, X. Huang, A. M. Jamison, M. Dredze, S. C. Quinn, J. W. Ayers, Twitter and Facebook posts about COVID-19 are less likely to spread misinformation compared to other health topics. *PLOS ONE* **17**, e0261768 (2022).
79. J. Kurkowski, "john-kurkowski/tldextract," *GitHub*, 2020; https://github.com/john-kurkowski/tldextract.
80. M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoli, A. L. Schmidt, P. Zola, F. Zollo, A. Scala, The COVID-19 social media infodemic. *Sci. Rep.* **10**, 16598 (2020).
81. L. Singh, L. Bode, C. Budak, K. Kawintiranon, C. Padden, E. Vraga, Understanding high- and low-quality URL sharing on COVID-19 twitter streams. *J. Comput. Soc. Sci.* **3**, 343–366 (2020).
82. M. R. DeVerna, F. Pierri, B. T. Truong, J. Bollenbacher, D. Axelrod, N. Loynes, C. Torres-Lugo, K.-C. Yang, F. Menczer, J. Bryden, CoVaxxy: A collection of English-language Twitter posts about COVID-19 vaccines, in *Proceedings of the International AAAI Conference on Web and Social Media* (2021), pp. 992–999.
83. K.-C. Yang, F. Pierri, P.-M. Hui, D. Axelrod, C. Torres-Lugo, J. Bryden, F. Menczer, The COVID-19 Infodemic: Twitter versus Facebook. *Big Data Soc.* **8**, 20539517211013860 (2021).
84. D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, The science of fake news. *Science* **359**, 1094–1096 (2018).
85. 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).
86. B. Golding, "Iffy index of unreliable sources," *Iffy.news*, 3 May 2020; https://iffy.news/index/.
87. N. F. Johnson, R. Leahy, N. J. Restrepo, N. Velasquez, M. Zheng, P. Manrique, P. Devkota, S. Wuchty, Hidden resilience and adaptive dynamics of the global online hate ecology. *Nature* **573**, 261–265 (2019).
88. T. M. J. Fruchterman, E. M. Reingold, Graph drawing by force-directed placement. *Softw. Pract. Exp.* **21**, 1129–1164 (1991).
89. J. W. Ayers, B. Chu, Z. Zhu, E. C. Leas, D. M. Smith, M. Dredze, D. A. Broniatowski, Spread of misinformation about face masks and COVID-19 by automated software on Facebook. *JAMA Intern. Med.* **181**, 1251–1253 (2021).
90. F. Giglietto, N. Righetti, L. Rossi, G. Marino, It takes a village to manipulate the media: Coordinated link sharing behavior during 2018 and 2019 Italian elections. *Inf. Commun. Soc.* **23**, 867–891 (2020).

91. D. Pacheco, P.-M. Hui, C. Torres-Lugo, B. T. Truong, A. Flammini, F. Menczer, Uncovering coordinated networks on social media: Methods and case studies. *ICWSM* **21**, 455–466 (2021).
92. L. Nizzoli, S. Tardelli, M. Avvenuti, S. Cresci, M. Tesconi, Coordinated behavior on social media in 2019 UK general election, in *Proceedings of the International AAAI Conference on Web and Social Media* (2021), pp. 443–454.
93. D. Weber, F. Neumann, Amplifying influence through coordinated behaviour in social networks. *Soc. Netw. Anal. Min.* **11**, 111 (2021).
94. M. Okada, K. Yamanishi, N. Masuda, Long-tailed distributions of inter-event times as mixtures of exponential distributions. *R. Soc. Open Sci.* **7**, 10.1098/rsos.191643, (2020).
95. D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003).
96. D. M. Blei, Probabilistic topic models. *Commun. ACM* **55**, 77–84 (2012).
97. A. K. McCallum, "Mallet: A machine learning for language toolkit" (2002); www.citeulike.org/group/3030/article/1062263.
98. H. M. Wallach, D. M. Mimno, A. McCallum, Rethinking LDA: Why priors matter, in *Advances in Neural Information Processing Systems*, Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, A. Culotta, Eds. (Curran Associates, Inc., 2009), vol. 22; https://proceedings.neurips.cc/paper_files/paper/2009/file/0d0871f0806eae32d30983b62252da50-Paper.pdf.
99. J. D. Hamilton, *Time Series Analysis* (Princeton University Press, 1st edition, 1994).
100. J. Durbin, S. J. Koopman, *Time Series Analysis by State Space Methods* (OUP Oxford, 2012).
101. N. Dias, G. Pennycook, D. G. Rand, Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media. *Harv. Kennedy Sch. Misinfo. Rev.* **1**, 10.37016/mr-2020-001, (2020).
102. S. Curry Jansen, B. Martin, The streisand effect and censorship backfire. *Int. J. Commun.* **9**, 656–671 (2015).
103. G. Pennycook, A. Bear, E. T. Collins, D. G. Rand, The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Manage. Sci.* **66**, 4944–4957 (2020).
104. C. A. Bail, L. P. Argyle, T. W. Brown, J. P. Bumpus, H. Chen, M. B. F. Hunzaker, J. Lee, M. Mann, F. Merhout, A. Volfovsky, Exposure to opposing views on social media can increase political polarization. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 9216–9221 (2018).
105. "Fighting coronavirus misinformation and disinformation," *Center for American Progress*, 18 August 2020; www.americanprogress.org/article/fighting-coronavirus-misinformation-disinformation/.

Acknowledgments: We thank R. Tromble and B. Nyhan for comments and feedback. **Funding:** This work was supported by the John S. and James L. Knight Foundation through the GW Institute for Data, Democracy, and Politics (to D.A.B. and L.C.A.), the National Science Foundation under grant no. 2229885 (to D.A.B.), and the National Science Foundation under grant no. 2029420 (to D.A.B.). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the John S. and James L. Knight Foundation. **Author contributions:** Conceptualization: D.A.B., J.G., J.R.S., and L.C.A. Methodology: D.A.B., J.G., A.M.J., J.R.S., and L.C.A. Investigation: D.A.B., J.G., J.R.S., and A.M.J. Visualization: D.A.B. and J.R.S. Funding acquisition: D.A.B. and L.C.A. Project administration: D.A.B. Supervision: D.A.B. Writing—original draft: D.A.B. Writing—review and editing: D.A.B., J.G., A.M.J., J.R.S., and L.C.A. **Competing interests:** D.A.B. has received consulting fees from Merck & Co. for participating in the 2021 and 2022 Merck Global Vaccine Confidence Expert Input Forums and has received a speaking honorarium from the United Nations Shot@Life Foundation. The other authors declare that they have no competing interests. The views expressed are those of the authors and do not reflect the official position of the U.S. Department of Health and Human Services, or the United States government. **Data and materials availability:** Data in this study were obtained from CrowdTangle for Academics and Researchers, a third-party data provider owned and operated by Facebook. CrowdTangle list IDs are provided in the references. Anyone with a CrowdTangle account may access these lists and the corresponding raw data. CrowdTangle's terms of service prohibit providing raw data to anyone outside of a CrowdTangle user's account. Researchers can sign up for a CrowdTangle account at https://help.crowdtangle.com/en/articles/4302208-crowdtangle-for-academics-and-researchers. All data needed to evaluate the conclusions in the paper are present in the paper and the Supplementary Materials or are available as processed data at Harvard Dataverse at this link: https://doi.org/10.7910/DVN/RLG5ED.

Submitted 16 February 2023
Accepted 7 August 2023
Published 15 September 2023
10.1126/sciadv.adh2132

ScienceAdvances

The efficacy of Facebook's vaccine misinformation policies and architecture during the COVID-19 pandemic

David A. Broniatowski, Joseph R. Simons, Jiayan Gu, Amelia M. Jamison, and Lorien C. Abroms

Sci. Adv., **9** (37), eadh2132.
DOI: 10.1126/sciadv.adh2132