Trust in vaccination is eroding, and attitudes about vaccination have become more polarized. To shed light on the evolution of social media discourse about vaccines in the context of the COVID pandemic, we conducted an observational study of Twitter vaccine discourse 75 days prior to and after the World Health Organization’s 11 March 2020 pandemic declaration. We report four main findings. First, the amount of vaccine discourse greatly increased over the course of the pandemic. Second, the two existing vaccine-focused camps were joined by a massive influx of partisan American political accounts. Antivaxxers solidified alliances with Republicans, and Public Health organisations with Democrats, increasing interaction and signal boosting. We also detect a smaller community of Unorthodox users who are more ambivalent about vaccines; this community moves towards the political right after the declaration. Finally, the evolution of both the moral and the non-moral language used by these groups follows pre-existing patterns of trust, suggesting a trust-first model of political engagement.

KEYWORDS
social network, COVID-19, coronavirus, trust, vaccine hesitancy
Trust in vaccination remains high but is eroding in many parts of the world (de Figueiredo et al., 2020). Decreased confidence in the safety, efficacy, and importance of vaccination may manifest as outright skepticism or sharing of conspiracy theories. It also manifests more subtly in vaccine hesitancy, which leads to questioning both of the need for vaccination and of the ulterior (usually financial) motives of doctors and the pharmaceutical industry. Hesitancy comes in degrees: some people are "accepters" while others are "fence-sitters" or "rejecters" (Rossen et al., 2019). Confidence is impacted by lack of information and access to misinformation (especially via ‘alternative medicine’ outlets). It is also undermined by distrust of medical and government sources, which are sometimes seen as overplaying the risk or severity of disease and underplaying adverse side effects of vaccines (Yaqub et al., 2014). It is difficult to assess how widespread these trends are, as the vast majority of studies of trust in vaccination have been conducted only in developed countries, especially in the United States, Western Europe, Australia, Japan, and Taiwan (Larson et al., 2018).

When people lose trust in medical experts and public health officials, they tend to turn to other sources, including family, friends, traditional and social media, and internet search and recommender systems. Online sources have come to the fore during pandemic lockdowns, including online sources that are optimized to capture and keep users’ attention, such as Facebook, YouTube, and Twitter — not to mention more recent entrants such as MeWe and Parler. What captures and keeps attention, however, needn’t be epistemically reliable. This development is concerning because exposure to negative opinions about vaccines on social media has been shown to be among the strongest predictors both of expressing such opinions oneself (Dunn et al., 2015) and of failure to vaccinate (Dunn et al., 2017).

A pair of recent, pre-COVID-19 studies found that the English-language discourse about vaccines on Twitter is highly polarized, that the anti-vaccine camp has greater reach and receptivity as measured by degree and PageRank, and that discussants tend to rely on and amplify just a few, non-independent sources (Sullivan et al., 2020a,b). It does not appear that this polarized network involves filter bubbles, in which participants are simply unaware of opposing views; instead, the network more closely resembles a pair of echo chambers, in which ingroup trust is high, outgroup awareness is also high, but outgroup distrust is high.

Generalizing from the pre-COVID-19 world to the post-COVID-19 world is inferentially fraught. Most of the discourse about vaccination revolves around well-established immunization to well-understood childhood diseases, seasonal influenza, and (more recently) human papillomavirus. With the exception of the seasonal influenza vaccine, these vaccinations tend to be administered to infants and children, whose caretakers must make proxy decisions about their health. Well-established vaccines may also suffer from the ‘curse of success,’ because their widespread administration has in many cases reduced incidence levels of the relevant diseases to near-zero.

COVID-19, by contrast, is not as well understood. Vaccines against it have (at the time of writing) only recently been approved, and administration remains an urgent but complex challenge.

To shed light on the evolution of social media discourse around vaccines in the context of the pandemic, we conducted an observational study of approximately three months’ of Twitter discourse — 75 days prior to the World Health Organization’s 11 March 2020 pandemic declaration through 75 days after the declaration.

---

1 In a recent philosophical analysis, Nguyen (forthcoming) argues that trust is best understood as an unquestioning attitude, so patients’ and parents’ questioning of the medical sector can be seen as an expression of distrust.
2 Social media sites are designed to sustain (and commoditize) users’ attention, as part of the ‘attention economy’. These are many cross-disciplinary studies that touch upon how social media sites optimize for engagement of users (contra other measures such as information veracity or epistemic well-being). See Alfano et al. (2018a, 2020) and Burr et al. (2018) for theoretical and empirical approaches to this idea.
3 For more on this way of distinguishing filter bubbles from echo chambers, see Nguyen (2020).
We report four main findings. First, the amount of vaccine discourse greatly increased. Second, authors in the vaccine discourse cluster into several large, interpretable groups. Third, much of the increase was due to the increased participation of accounts associated with partisan American politics: Republicans and Trump supporters flooded the conversation and were allied with Antivaxxers while Democrats joined the fray and were allied with health organizations. We also detect a smaller community who are more ambivalent about vaccines and seem to play a bridging role between the anti-vax right and the pro-vax center-left. Fourth, the manner in which these groups talked about the pandemic evolved in a way that followed pre-existing patterns of trust, rather than trust being dictated by language use. In other words, Antivaxxers and Republicans were already socially connected before the pandemic struck, as were Democrats and Public Health; after the pandemic was declared, the way these groups used language (including moral language) converged with their preexisting social connections. This suggests a “trust-first” rather than an “evidence-first” or “values-first” model of the evolution of vaccine discourse. We conclude by discussing the implications of this research, which suggests that political partisans have the power — if not to persuade — then both to set the agenda and to frame seemingly non-partisan public health issues.

2 | METHODOLOGY AND DATA DESCRIPTION

In this section, we explain the methods used to collect, clean, and curate our dataset.

2.1 | Data collection

We queried the Twitter Streaming API with a series of vaccination-related keywords, hashtags and short expressions between December 2019 and June 2020 (see Supplementary Material A.1). Taking as a reference point the date the World Health Organization declared COVID-19 a pandemic (11 March 2020), we divided our dataset into a symmetric timespan of tweets comprised of approximately 1.3 million original tweets and 18 million retweets between December 27th and May 26th (see SM A.2).

A tweet can either contain original content, be a pure retweet of another tweet, or be a quote tweet which combines a retweet with additional commentary. We considered only pure retweets in our analysis. Retweets generally signal endorsement of content and an attempt to signal-boost. Anecdotal evidence from the early 2010s, even before Twitter began to formalize the concept of retweeting in its application interface, found that retweets serve three purposes: “spreading tweets, to start a conversation, and to draw attention to the originating user” (Cheong, 2013; Boyd et al., 2010).

We also collected information about the retweeted content. What counts as retweeted content can be interpretively subtle. If an original tweet is retweeted, then the content is the original tweet and the retweeted author is the author of that content. If a quote tweet (or a series of quote tweets) is retweeted, then we considered the retweeted content and author to be that of the most recent comment, not the original tweet or any quote tweet lower in the chain. This preserves the endorsement flavor of retweets: if $x$ says something, $y$ quotes to disagree, and $z$ retweets $y$’s disagreement, it is likely that $z$ also disagrees with $x$. For convenience we will talk of tweets and tweet contents below, but it should be noted that these can sometimes be quote tweets. Finally, note that the Twitter API functions in such a way that intermediate retweets are not stored: if $y$ retweeted $x$’s tweet $T$, and $z$ retweets $y$’s retweet, the

---

4See URL = [https://www.who.int/emergencies/diseases/novel-coronavirus-2019/interactive-timeline/](https://www.who.int/emergencies/diseases/novel-coronavirus-2019/interactive-timeline/) results do not corroborate the findings of Chan (2020); Rossen et al. (2019) or Amin et al. (2017), nor are they consistent with the hypotheses, noted above, that Republicans would score especially high on the binding foundations and demonstrate the largest increase in the sanctity foundation.
data will show only a retweet of $x$ by $z$.

2.2 | Network Construction

We generated a retweet network, a weighted directed network where nodes are authors and the weight of an edge from node $u$ to node $v$ represents the number of times that user $v$ retweeted user $u$. Self-retweeting was discarded. Users that only retweeted but never authored a tweet were discarded. The intended interpretation of the network is that if $v$ retweeted $u$, this means that $u$ influenced $v$. Retweet networks have been used before to study community engagement and the spread of fake news [Bovet and Makse, 2019; Sullivan et al., 2020a,b].

We pruned the network by only considering the principal weakly connected component, thus excluding clusters isolated from the main conversation. The resulting network has $\sim 380K$ nodes and $\sim 3.6M$ edges. Both weighted and unweighted degree versions of the network follow a statistically significant ($p = 0.0127$ and 0.0092 respectively) power law distribution when compared with the lognormal distribution [Alstott et al., 2014]. In particular, 80% of nodes had fewer than five neighbors and only 5% of nodes were retweeted by more than 20 distinct nodes, which shows that a small minority was highly influential. Power law distributions are recurrent in social media [Cheong, 2013] and in particular for retweet networks [Lu et al., 2014], which suggests that our vaccine-focused subgraph is representative.

2.3 | Community detection and characterization

In order to identify the different communities, we used Gephi’s implementation of the Louvain modularity algorithm developed by Blondel et al. [2008]. The modularity of a network measures the strength with which a network can be divided into groups. The measure works by computing the fraction of edges that fall within the given communities minus the expected fraction if edges were distributed at random, but keeping the same (weighted) degree distribution. Therefore, a high modularity indicates that members of the communities are unexpectedly bound to each other, holding node centrality fixed and having randomness as a baseline. Gephi’s implementation can handle weighted and directed networks like the retweet network we generated, and allows for different resolutions as developed by Lambiotte et al. [2009].

We implemented the community detection algorithm including randomization, using edge weight and with a resolution value of 1. We obtained a network modularity score of 0.608 and the implementation found 231 communities, with the top five largest comprising 80% of the population and the top eight largest comprising 95%. Manual inspection showed that similar community partitions emerged in repeated implementations of the algorithm. To characterize the communities we considered the top verified accounts in each cluster, as well as the typical hashtags used by the communities.

For each community we also calculated the Gini coefficient [Gini, 1909] to measure (in)equality in the distribution of influence. The Gini coefficient is typically used to compare wealth distributions across countries, but it has recently been used to analyse structures of social networks. The Gini coefficient is a real number from 0.0 to 1.0 inclusive, where 0.0 indicates a perfectly equal distribution of resources among all members of a group and 1.0 indicates the complete monopolization of wealth by just one member of the community [Kelly, 2012].

For our retweet networks, we define a unit of ‘wealth’ as the influence a Twitter user, $x$, can exert over another, $y$, such that $y$ is disposed to retweet a post by $x$; i.e. $x$’s total ‘wealth’ is the number of their posts that are retweeted by others (including duplicates). We calculated the Gini coefficient for each community\(^5\) both pre- and post-declaration.

---

2.4 Corpus-based analysis of retweets

To analyse the corpora of tweets per group, we employed a word-counting approach modeled on Linguistic Inquiry and Word Count (LIWC), which was developed by Pennebaker et al. (2015). Over the last few decades, Pennebaker and his collaborators have shown that this seemingly simple method can reveal quite a bit about individuals, their relationships, and the groups to which they belong (Pennebaker, 2011). In the interest of open science, we used the R package LIWClike (Benoit et al., 2018), which imitates and expands the functionality of LIWC. As is standard in LIWC analyses, we did not remove stopwords or perform any stemming/lemmatizing of the corpora.

An advantage of the LIWC approach is the ability to create and share custom dictionaries for categories of interest. Recent interdisciplinary work has shown that it is possible to extract a moral signal from natural language using various tools (Alfano et al., 2018b). To better characterize the communities described above, we employed two custom LIWC dictionaries that are keyed to various moral concerns. The Moral Foundations Dictionary (MFD) measures the number of words in a text associated with care (versus harm), fairness (versus unfairness), authority (versus insubordination), loyalty (versus disloyalty), and sanctity (versus corruption). These domains or "foundations" feature in Moral Foundations Theory (Graham et al., 2013) and are conceived of as topics towards which individuals are differentially sensitive. MFD includes two sub-dictionaries (one related to virtues, the other to vices) for each foundation. We also passed the corpora through an alternative dictionary, the Morality-as-Cooperation Dictionary (MAC-D), which is based on the Morality-as-Cooperation hypothesis (Curry et al., 2019a,b). According to this theory, morality is best understood functionally, as a set of rules, norms, sentiments, institutions, and technologies that help us to solve recurrent problems of cooperation. Morality-as-Cooperation posits seven domains or elements (family, group, reciprocity, heroism, deference, fairness, and property), and MAC-D contains one sub-dictionary for each, covering only virtues and not vices.

3 RESULTS

In this section, we report our results, beginning with community characterization, then moving to descriptive statistics, network evolution, linguistic evolution, and corpus analysis.

3.1 Community characterization

For this study, we decided to focus on the top five communities. Not only are these clearly interpretable (Figure 1), but they contain ~ 80% of the nodes and are responsible for ~ 90% of the retweets. The top community we labeled 'Democrats', and it includes Democratic politicians and some center-left media. The second largest community we labeled 'Republicans'; the third 'Unorthodox'; the fourth 'Public Health'; and the fifth 'Antivaxxers'. Table 1 summarizes the statistics for the communities.

---

6 Details of these dictionaries are available at URL = HTTPS://OSF.IO/EZN37/ (MFD) and URL = HTTPS://OSF.IO/86QRY/?VIEW_ONLY=CD9E0EA3659422DB052AB3359560A20/ (MAC-D). The usage of LIWC dictionaries is often preferable to more complex techniques (such as deep neural networks) as they are more explainable and interpretable by the end user. LIWC dictionaries are in a simple plain text format, and the inner workings of the algorithm are well-documented and can be replicated by hand if needed.

7 For more information about the remaining of communities, please see the Supplementary Material.
**TABLE 1** Summary statistics for the five communities analyzed in this study. To protect the privacy of individual Twitter users, only public figures (with a 'verified account' badge by Twitter and/or notable enough to have their own Wikipedia page) and suspended/deleted accounts are listed as examples within. (i) % of retweets: These account for both retweeting and being retweeted within the network. (ii) Popular Hashtags: All of these are within the top 15 hashtags that each community used, after preprocessing and using a term frequency - inverse document frequency (tf-idf) representation to identify salient content for each community. (iii) Representative Nodes: The users marked with asterisks (*) have either been suspended by Twitter or have been deleted, at time of writing. The typical reason for suspension is violation of the Twitter Terms of Service.

<table>
<thead>
<tr>
<th>Community name</th>
<th>Proportion of nodes</th>
<th>% of retweets</th>
<th>Popular Hashtags</th>
<th>Representative Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democrats</td>
<td>~ 24%</td>
<td>~ 20%</td>
<td>#moronpresident, #trumpslump, #gopvirus, #trumpgenocidedfor-profit, #trumpburialpits</td>
<td>JoeBiden, Kamala-Harris, SenWarren, BillGates, CNN, ny-times, washingtonpost, ABC, businessinsider, MSNBC, guardian (theguardian), TIME, BBCWorld</td>
</tr>
<tr>
<td>Republicans</td>
<td>~ 18%</td>
<td>~ 35%</td>
<td>#kungfu, #notest, #boycottchina, #trumpcovidgate, #illuminati</td>
<td>realDonaldTrump, mikepence, RealCan-daceO, WhiteHouse, FoxNews</td>
</tr>
<tr>
<td>Unorthodox</td>
<td>~ 16%</td>
<td>~ 6%</td>
<td>#youthsoncoronake, #listentotelexperts, #noisundays, #locksouthafricadown, #coronavirusghana</td>
<td>BernieSanders, Trevornoah, specta-torindex, jacobinmag, NaomiAKlein, BBC-Africa, DrTedros, News24 (African Media)</td>
</tr>
<tr>
<td>Public Health</td>
<td>~ 13%</td>
<td>~ 7%</td>
<td>#epidemic, #hepati-tisa, #immunoonc, #whatwedoingphar-macy, #scteenvax</td>
<td>CDCgov, WHO, EU_Health, Unio-fOxford, UNICEF, ProfPCDoherty (Nobel Laureate Immunology), VaccinesToday, CDCFlu, newscientist, CEPIvaccines, gavi</td>
</tr>
<tr>
<td>Antivaxxers</td>
<td>~ 8%</td>
<td>~ 22%</td>
<td>#illuminati, #notest, #mykidsmy-choice, #vacciner-oulette</td>
<td>stopvaccinating*, StopVaxTyranny*, EpigeneticWhisp*, vaxxplained*, Just-SayNo2Vax*, va_shiva, Jimcorrsays</td>
</tr>
</tbody>
</table>
**3.2 Descriptive statistics**

The graphs in Figure 2 give a basic idea of how activity changed before and after the declaration for each of the communities that we studied. All of them increased in size (author count), number of tweets they generated (tweet count), number of times they were retweeted, and number of times that they retweeted. More saliently, Republicans increased their tweeting and retweeting activity in an even greater proportion than other groups.

In addition to differences between groups in total numbers of retweets, there were also substantial variations in retweeted group. Table 2 shows the change in absolute numbers and ratio of post- to pre-declaration tweets.
Table 2 shows that, unsurprisingly, the largest increases were in groups signal-boosting their own members. This depended in part on how active the groups were pre-pandemic: Antivaxxers and Public Health increased at the lowest rate, while Republicans retweeted Republicans at a vastly higher rate. There were also other notable increases: Antivaxxers also signal-boosted Republicans at a much higher rate than vice-versa. The asymmetric rate increase was less notable; however, because Republicans made such a large contribution to the dataset, the increased number of retweets of Antivaxxers made by Republicans was still quite high. Also of note was the increase in retweets of both Antivaxxers and Republicans by the Unorthodox. While not a large increase in absolute numbers, it made for a substantial increase in ratio.

4 NETWORK EVOLUTION

Although the retweeting activity between communities changed over time, the network itself was polarised throughout. As Figures 3a and 3b show, Democrats and Public Health form one pole of the polarised network and Antivaxxers...
and Republicans the other. The Unorthodox group was the only one with a noticeable topological shift, moving to occupy an intermediate position between the two poles. To quantify this impression, we used hierarchical clustering on a distance measure of group retweets (See SM A.3), represented as dendrograms in Figures 4a and 4b. This confirms the overall picture of polarisation, along with the shifting position of Unorthodox.

The results for the Gini coefficient were surprising (we summarise key findings below, with full results in the Supplementary Materials). We expected that the Gini coefficients for the various communities would tend to increase, as that would reflect the well-studied phenomenon of preferential attachment, i.e. rich-get-richer dynamics over time [Barabási and Albert 1999; Cheong 2013]. However, our empirical results indicate that in every single community, the Gini coefficient decreased from pre- to post-declaration time slices. The unexpected decline in Gini coefficients could be explained by two mutually consistent hypotheses. First, perhaps there was a ceiling effect as the pandemic began to take its toll and people looked for new sources of information. If trust was already directed towards the most prominent nodes in the network, then these established sources would perforce become less prominent because they had already gained as many followers as they could. Second, there may have been an increase in conspiracy theorizing (as the prominence of the ‘illuminati’ hashtag in the Republicans and Antivaxxers suggests); if people began to distrust official sources, that could lead them to follow and retweet what had previously been seen as more fringe opinions.

4.1 Linguistic Evolution

There were 10 total corpora (5 communities x pre- and post-declaration). The corpus associated with each group more than doubled in size from pre- to post-pandemic declaration. The smallest initial corpus was associated with Unorthodox, which increased from 323,730 words to 1,882,664 words. The largest initial corpus was Antivaxxers, which increased from 1,394,630 words to 2,664,886 words. However, the largest increase was in Republicans, which went from 533,620 words to 4,509,393 words, suggesting a surge in interest in a topic that previously had been of
relatively little concern to these users.

As with the network graph, we used hierarchical clustering to show similarities between the language used by groups before and after the pandemic declaration (see SM A.3 for specifics).

As Figure 5 suggests, before the pandemic declaration, there were two thematically-unified discourses: one about politics, carried out by participants from across the political spectrum from Republicans to Democrats to Unorthodox, and another about health, carried out by Public Health and Antivaxxers. Of course, this doesn’t mean that Democrats and Republicans agreed: rather, they were still talking about the issues in a broadly similar way. Likewise, Public
Health and Antivaxxers didn’t agree, but they were using much the same language. After the pandemic declaration, however, we observe a reshuffling of the discourses: Democrats and Public Health start using the same words, while Republicans, Antivaxxers, and to some extent the Unorthodox begin using the same words. The main difference between alternative methods of clustering is where they place the Unorthodox category: most have them shifting to nearer the Republicans/Antivaxxers pole, but the custom moral dictionaries cluster them closer to the Democrats (Figures 6 and 7).

**FIGURE 6** Dendrogram for LIWC-specific content columns.

**FIGURE 7** Dendrogram for custom MFT/MAC virtue dictionaries.

Notably, we observe that the patterns in both the moral and the non-moral discourse evolved to match the preexisting pattern of social connection. Initially, the moral and non-moral language used by Antivaxxers and Public Health were most similar to one another, and the moral and non-moral language used by Republicans, Democrats,
and Unorthodox were most similar to one another. However, after the pandemic declaration we see Republicans and Antivaxxers expressing the same moral and non-moral concerns, Democrats and Public Health expressing the same moral and non-moral concerns, and Unorthodox again playing an ambivalent role in between the two polarities.

4.2 | Corpus Analysis

4.2.1 | Standard LIWC

Using LIWC dictionaries, we can begin to shed some light on exactly how the language of these different communities shifted. (The full table is here or available upon request.)

**FIGURE 8** Changes in score on selected LIWC components pre- and post-declaration. Components on X axis show top 15 relative score changes. Y axis shows percentage of max pre/post score across all groups (for raw scores, see SM B.1). Endpoints show pre- and post-declaration values, and arrow indicates direction of change.
As Figure 8 shows, changes in individual groups also shed some light on their convergence. Most notable is the dramatic increase, for Republicans and Antivaxxers, of discourse around ‘Home’, ‘Masculinity’, and ‘Money’. This, combined with the decrease in ‘Female’/‘Family’ talk, suggests a strong reorientation of vaccine discourse away from issues of motherhood and towards traditionally masculine issues such as personal liberty and economic growth. By contrast, Democrats and Public Health showed only small changes on these issues, and Unorthodox tweeters showed increases but not to the same level. Democrats and Public health officials also show decreases in discourse around ‘Anger’, ‘Bodies’, and the ‘Feel’ category, suggesting a shift towards a more neutral, dispassionate mode of discussion. Finally, there are areas of convergence: Religious discourse remained low and relatively unchanged for Democrats and Public Health officials. It increased for Republicans and Unorthodox tweeters, while Antivaxxer Religious discourse decreased, bringing about more of a convergence with the Republican/Unorthodox camp.

4.2.2 Custom moral dictionaries (MFD and MAC-D)

In previous work, Haidt and colleagues have found that more politically conservative people and communities tend to place greater emphasis than liberals or leftists do on the so-called “binding” foundations of loyalty, authority, and sanctity [Graham et al. 2009]. In addition, political conservatives are typically found to be higher in dispositional disgust-sensitivity, which is associated with the sanctity foundation [Inbar, Pizarro and Bloom 2009] though see [Elad-Strenger, Proch and Kessler 2020] for a dissenting view). This should lead us to expect that the Republicans corpus would score especially high on the three binding foundations both pre- and post-pandemic declaration, and that the Republican corpus would exhibit a more-pronounced increase in language related to sanctity from pre- to post- than other corpora.

When it comes to moral foundations and COVID-19 in particular, it has been hypothesized that normative health behaviors during the pandemic may be partially explained by individual differences in moral foundations. In particular, [Chan 2020] provides evidence that Americans who score high on the care and fairness domains are more likely to report staying at home, support wearing facemasks, and respect social distancing, while those who score high on the sanctity domain are more likely to report wearing facemasks and comply with social distancing, yet refuse to stay at home. The former case is potentially due to a general recognition of collective responsibility and a shared duty of care to others; the latter concerns COVID spread as an issue of avoiding contamination (see also Amin, Bednarczyk, Ray, Melchiori, Graham, Huntsinger and Omer 2017).

In addition, differential attunement to moral foundations may drive vaccine hesitancy. In particular, [Amin et al. 2017] offer evidence that parents who score high on the sanctity foundation are especially likely to be fence-sitters or rejecters when it comes to vaccinating their children, and that the care and fairness foundation are unrelated to vaccine hesitancy. Along the same lines [Rossen et al. 2019] find that parents high on the sanctity foundation, low on the authority foundation, and high on the care foundation are more likely to be rejecters.

Finally, because Antivaxxers tend to be primarily worried about the health of their families, especially their own children, we should expect that this community would score especially high on the Family element of MAC-D. And because it is part of the job description of public health officials to be concerned for the whole group rather than only individuals to whom they have special ties, we expect that this community would score especially high on the group solidarity element.

Using the custom moral dictionaries, we can test these hypotheses and shed further light on exactly how the language of these different communities shifted. Once again, the full table is here or available upon request. We here discuss selected columns in Figure 9.
Moving along the five moral foundations from pre- to post-declaration, we can pinpoint two distinct patterns that indicate the consolidation of an Antivax-Republican alliance at one end, and, a Democrat-Public Health partnership at the other. Taken together, these two trends evidence a willingness among but not across both sets of communities to adjust their initial moral emphasis so as to accommodate the values of their closest interlocutors. Interestingly, these dynamics of moral (dis)agreement depart somewhat from traditional models of (de)polarization. While it holds true that communities situated at opposing ends of the spectrum move further apart; those who initially emphasize just a few shared foundations seem remarkably ready to negotiate their remaining moral differences. Thus, while polarization between poles persists, within each pole, depolarization takes place. On the one hand, this pattern suggests that moral compromise is possible; on the other, it also suggests that local compromise with your closest allies can lead to increased global disagreement (Uslaner, 2002). We canvas one possible explanation for this observation in §5.4.

For now, it bears spelling out that all groups demonstrate a decrease in the proportion of care-related words, with the largest decrease among Antivaxxers who end up converging with Republicans in placing less emphasis on care. There was relatively little change in the use of fairness-related words, though it is noteworthy that Republicans and Unorthodox show an increase while the other groups have mixed results or (in the case of Antivaxxers) a decrease, again indicating convergence on moral values. Likewise, on the loyalty foundation, Democrats and Public Health converge in placing additional weight on the virtue and little emphasis on the vice, while Republicans and Antivaxxers move in the opposite direction and place less emphasis on the virtue and more on the vice. Republicans and Antivaxxers also see a relatively large uptick in emphasis on the vices of authority. When it comes to the sanctity dimension, the virtue and vice dictionaries tend to move in opposite directions. Only the Public Health community increases for both virtue and vice. These results do not corroborate the findings of Chan (2020), Rossen et al. (2019), or Amin et al. (2017), nor are they consistent with the hypotheses, noted above, that Republicans would score especially high on the binding foundations and demonstrate the largest increase in the sanctity foundation.

Next, we consider the results for MAC-D in Figure 10.
decrease in the Antivaxxers community. This is consistent with the results for the built-in LIWC ‘Family’ dictionary mentioned in the previous section and our suggestion that changes in the discourse reflect a reshuffle and politicization of the conversation, with Republicans and Antivaxxers mutually exploiting one another while Democrats made common cause with Public Health. As mentioned above, we hypothesized that Antivaxxers would score especially high on the family element of MAC-D. This is true of Antivaxxers pre-declaration, but they differ much less from the other groups post-declaration. We also hypothesized that Public Health would score especially high on the group element. This is true and becomes more pronounced from pre- to post-declaration.

We also observe that group solidarity increases for all groups except Republicans, reciprocity increases for all communities, heroism increases for all but Public Health, and deference increases for all but Democrats. Fairness is more of a mixed bag, though it is noteworthy that Public Health stands out as showing by far the most emphasis on this element post-declaration. This may be due to the fact that the fairness element in MAC-D specifically has to do with fair (typically equal) distribution of resources, which would also be the best public health strategy for containing the pandemic and distributing vaccines for COVID-19. Finally, on the property dictionary, we observe Republicans moving down while Antivaxxers move up to match them, again suggesting that moral values evolved based on preexisting connections of trust.

5 | DISCUSSION

In this section, we discuss the results canvassed above, then turn to limitations of the current study and directions for future work.

5.1 | Network communities and corpus analysis

Our approach employed mixed methods, using modularity to detect communities on the social network of retweets, then supplementing that information with corpus analyses using pre-built and custom LIWC dictionaries. Previous work on Reddit has demonstrated the utility of mixed methods for understanding the evolution of online engagement with conspiracy theorising [Klein et al., 2019]. The fact that these analyses largely converge in identifying five communities that evolved both in their social connections and in their language use corroborates our main findings. Republicans and Antivaxxers were already connected socially before the pandemic was declared, and they became more connected and started using both moral and non-moral words in the same way more frequently after the declaration. Their concerns seem to relate primarily to masculinity and money, with a large drop-off in their focus on family
and femininity. By contrast, Democrats and Public Health were already connected socially before the pandemic was declared, and they became more connected and started using words in the same way more frequently after the declaration. The Unorthodox group is exceptional in that they play an interstitial role socially and seem to have been pulled partially into the orbit of the Republicans/Antivaxxers pole of the network over time.

5.2 Emergent behavior: from signal-boosting to agenda setting, priming, and framing

The phenomenon of “signal-boosting” members of one’s own group is familiar to anyone who has spent time on Twitter: people tend to retweet messages published by those they view as co-partisans or allies (Heilig, 2015). Our findings in Sections 3.1 and 3.2 are consistent with this impression. However, we also observe interesting dynamics in which groups see each other as allies, leading to cross-group retweets. Here are three possible models of how groups determine who counts as an ally:

- **Evidence-first dynamics**: Treat a group as an ally worthy of having their signal boosted if and only if they tend to publish objectively reliable information.
- **Values-first dynamics**: Treat a group as an ally worthy of having their signal boosted if and only if they tend to publish information that expresses values your group shares.
- **Trust-first dynamics**: Treat a group as an ally worthy of having their signal boosted if and only if you already trust them, regardless of the reliability and value-ladenness of the information they tend to publish.

According to the evidence-first model, people are constantly on the lookout for reliable sources, and they update their rosters of friends and adversaries (people to trust and people not to trust) based on whether others offer reliable testimony (Craig, 1999; Mercier, 2020). According to the values-first model, people are instead disposed to trust others who share their moral outlook, even if those others are systematically unreliable in their testimony. This makes a certain amount of sense, as most theories of trust posit that someone is trustworthy only if they bear the trustor goodwill or have some related motivational disposition (e.g., Baier, 1986; Holton, 1994; Jones, 2012). Finally, according to the trust-first model, people first latch onto others whom they view as trustworthy, then follow them somewhat blindly, even to the point of being willing to accept seemingly unreliable information and to update their own values to match those of the trustee (Levy and Alfano, 2019).

Our findings are most consistent with the trust-first model. Not all communities deepened their ties to or began expressing themselves (proportionally) more like Public Health, even though Public Health was clearly the most reliable testimonial community in this discourse. This dynamic suggests that the evidence-first model is mistaken. Next, several groups significantly modified the values that they expressed in their language, most notably the Antivaxxers. Antivaxxers seem to have done so in order to match the values of their closest interlocutors, the Republicans. What appears to have happened is that Republicans flooded the discourse after the pandemic was declared. They were already socially connected to the Antivaxxers and Unorthodox communities, and those connections became stronger over time. Moreover, the Antivaxxers and Unorthodox communities shifted the types of moral and non-moral language they used to better match the language used by Republicans. At the same time, Democrats and Public Health already had ongoing ties of trust, and these ties deepened over the course of the pandemic. The Unorthodox community seems to have been pulled somewhat in the direction of the invading Republicans but not as far as the Antivaxxers were.

But what is the function of this brigading of the vaccine discourse by Republicans? Communications theory suggests that there are three main mechanisms of the media, including social media: agenda setting, priming, and
framing (McCombs and Shaw, 1972; Iyengar and Kinder, 1987; Scheufele and Tewksbury, 2007). Agenda setting refers to which questions and inquiries are on people’s minds. For instance, “What should we do about COVID-19?” Priming refers to the standards thought to be relevant to these questions. For instance, one might think that whether a new vaccine has passed muster via a peer-reviewed phase 3 trial is enough to settle whether it is safe and effective, or one might think that whether Donald Trump tells you it’s safe and effective is enough to settle the question. Finally, framing refers to the context within which a question or topic is thought to reside. For instance, one might think of vaccination as a matter of public health, or one might think of it as a matter of economics and partisan politics.

Our results relate to all three of these dynamics in a way reminiscent of Benkler et al.’s (2020) results showing concerted efforts by Republicans to set the agenda, prime standards, and frame questions to promote their own conspiratorial narratives during the runup to the 2016 US Presidential election. First, Republicans seem to have set the agenda for a large part of the discourse by repeatedly raising questions about the necessity for and the safety and effectiveness of vaccines. The fact that they made common cause with Antivaxxers is consistent with this, as is the fact that the discourse became less about health, risk, and death. Second, Republicans may have moved the goalposts in terms of what counts as evidence for the claims under discussion by making the discourse more about money than directly-relevant topics like health.

5.3 Limitations and future work

In this section, we review the limitations of our study, as well as opportunities for further exploration. Firstly, there are limitations to the dataset and corpora used in our analysis. Despite the richness of data made available to us, we are unable to capture the entirety of the global Twitter discourse on vaccines, as Twitter only makes roughly 1% of all tweets available to researchers without a commercial contract. Our corpus is also mostly in the English language, which represents roughly 68% of Twitter chatter worldwide pertaining to COVID. Also, as the pandemic is still ongoing as of time of writing, further dynamics would be interesting to follow as time progresses, for instance as vaccines get deployed at scale and a handful of adverse events inevitably occur. Future work to address these issues may include augmenting our corpus with other multi-lingual, and more recent, COVID Twitter datasets. Recent work on estimating the distribution of missing tweets from the API stream may also offer a more nuanced picture of any gaps.

Secondly, there are limitations to the natural language processing techniques used. LIWC dictionaries have the advantage of transparency and ease of understanding, but they may not be as advanced as other techniques, such as sentiment analysis, word2vec/doc2vec, and modern deep learning language models (such as GPT-3 and BERT). Furthermore, it is hard to assess the reliability of the signals obtained using MFD and MAC-D dictionaries. In particular, we noted that the results using MFD do not match the expectation that politically conservative Republicans would place a greater emphasis on ‘binding’ foundations of loyalty, authority, and sanctity. But this might be due to the effect that COVID-19 had on Republicans’ language. Our findings may indicate that there are problems with MFD, at least in the context of the COVID-19 pandemic. Perhaps the binding and more specifically the sanctity foundations do not work as advertised during a public health emergency. Alternatively, our findings may indicate a deeper problem with Moral Foundations Theory itself. Indeed, despite enjoying widespread adoption within academia and even crossing the mainstream into popular culture, Moral Foundations Theory has also attracted sustained criticism (Suhler and Churchland, 2011; Gray and Keeney, 2015).

Also of potential interest is the use of sentiment analysis on tweets, which is a powerful heuristic to measure the general ‘feel’ of social media discourse. However, because extant applications include, say, analysis of individual movie

8See summary statistics according to other dataset curators: see e.g. URL = **[https://github.com/echen102/COVID-19-TweetIDs](https://github.com/echen102/COVID-19-TweetIDs)**.
reviews and user sentiment pertaining to a product, these cannot be directly applied to our corpus because our data deals with health-related sentiment (pro-vaccination versus anti-vaccination), as they textually differ from traditional conceptions of sentimental polarity (positive/like versus negative/dislike). Furthermore, the short lengths of tweets are sub-optimal for extant methods. Hence, future work in this domain might include training classifiers using the method of [Shah et al., 2019]. Finally, other metadata sources attached to tweets may be able to enrich such work. Improvements in this regard include utilizing the geolocation data of users, co-tweet networks (similar to co-citation networks), LIWC analysis of user properties such as 'description', using the 'verified' badge of a user as a proxy for epistemic status, and analyzing past behavior and tweet history of users.

5.4 Conclusion

As with many social network studies, our study focuses on the network properties of the communities engaging in Twitter vaccination discourse during a relatively brief time window. Yet we observed a legacy network in which communities had already formed into clusters and made connections with other communities. The trust-first hypothesis raises the further question of why Republicans and Antivaxxers were already socially connected well before the pandemic struck.

One possibility, suggested in a recent popular article by DiResta & Lotan [2015] (see also DiResta 2020), is that Twitter's content moderation aimed at medical misinformation led to a shift in Antivaxxer's approach. Rather than straightforward medical misinformation (e.g., "Vaccines cause autism."), Antivaxxers began to seek political cover by allying with Republicans and reframing their message in political terms (e.g., "Mandatory vaccination is tyranny.").

This tantalizing hypothesis is consistent with our finding that Antivaxxers and Republicans reacted to the pandemic declaration by placing greater emphasis on the vices related to loyalty and authority. Silicon Valley has been more hesitant to moderate political speech, which makes politically-motivated misinformation harder to police. The possibility that selective content moderation might shift the form in which misinformation is expressed deserves careful future research.

Finally, while DiResta & Lotan's hypothesis plausibly explains moral depolarization between Antivaxxers and Republicans, it is silent on why we see a similar kind of value-assimilation among Democrats and Public Health. Contrary to being moderated on account of spreading medical misinformation, Public Health institutions aim to publish accurate content. Consequently, albeit more speculatively, these institutions neither need nor are motivated to bootstrap their agenda by adopting the frame of a political ally. In keeping with these observations, it is worth emphasizing that Democrats and Public Health do not converge to the same extent as Antivaxxers and Republicans (see §4.2.2). Though they do move toward each other along several foundations, the distance between them decreases chiefly in a comparative sense: it is in part because Antivaxxers and Republicans converge so strongly that the links between Democrats and Public Health become more pronounced. This is not to say that Democrats and Public Health are entirely insensitive to each other's values; however, it does suggest that their mutual moral sensitivity is sharpened by perceived consolidation at the opposite end of the spectrum. If this way of thinking is on the right track, future research could test whether the observed trends of (de)polarization exhibit an 'enemy of my enemy' dynamic. However dormant in academic circles, this approach intuitively makes sense of both alliance consolidation and community polarization.

In this paper, we showed how the online conversation about vaccines was barnstormed by political partisans, especially Republicans, once the COVID-19 pandemic was declared by the World Health Organization. Using both social network analysis and partially-automated corpus analysis, we demonstrated that Antivaxxers and Public Health were initially in dialogue with one another, but that Antivaxxers were also socially connected to Republicans. After the
pandemic was declared, all groups increased their engagement, but Republicans increased the most and seem to have influenced their closest interlocutors (Antivaxxers, followed by Unorthodox) to engage with them and to update the language they used to talk about vaccination — including both moral and non-moral language. At the same time, Public Health and Democrats grew closer together both socially and semantically. This pattern of results suggests that the dynamics of online discourse are driven by social connection and trust, rather than by deep, underlying, static values.
references


Biographies redacted for blind review.
We queried the Twitter Streaming API with a series of vaccination-related keywords, hashtags and short expressions. The full list of queries is as follows (including both English and some Dutch): 'vax', 'vaxxed', 'vaccine', 'vaccination', 'vaccinations', 'vaxsafe', 'vaccines work', 'vaccines revealed', 'vaccinesrevealed', 'novax', 'no vax', 'no-vax', 'antivax', 'anti-vax', 'immunisation', 'Vaccin', 'Vaccinations', 'vaccine injury', 'vax injury', 'vaccinatieschade', '#vax', '#vaxxed', '#vaccine', '#vaccination', '#vaccinations', '#vaxsafe', '#vaccineswork', '#vaccinesrevealed', '#novax', '#antivax', '#immunisation', '#Vaccin', '#Vaccinations', '#vaccineinjury', '#vaxinjury', '#rvp', '#rijksvaccinatieprogramma', '#vaccineinjury', '#anti-vax'.

To check for the presence of bots, we examined the top 50 authors by post and by retweet count. Upon manual inspection, we found that most prominent accounts do not behave like bots. Furthermore, only users that generated original tweets or quote tweets were taken into account; users that only retweeted were disregarded from our analysis and from the previous statistics.

The former amounts to sharing the same content another user posted, and the Twitter API functions in such a way that intermediate users are not stored - if y retweeted x's tweet T, and z retweets y's retweet, we will only get edges from x to y and from x to z, and nothing from y to z.

We also performed power law analysis on (a) tweet decay (i.e., the number of hours it takes a tweet to stop being retweeted), (b) retweet count per tweet, and (c) retweet count per author. Power law distributions are recurrent in social media and in particular Twitter \cite{Lu2014}. They reveal the fact that social influence is a highly centralized phenomenon, with a few authors and tweets garnering most of the attention. Only 5% of tweets continued to be retweeted after four days. Similarly, only 5% of tweets were retweeted more than 28 times, and only 5% of authors were retweeted more than 47 times. In fact, in all three cases we found that the empirical data follows a power law distribution with statistical significance (via comparison to lognormal distribution; c.f. \cite{Alstott2014}).

Given the power law distribution, we collected retweets of tweets made within the timeframe so long as they were made within 4 days. Hence to fall into the pre-declaration dataset, a tweet had to occur between December 27th and March 11th (74 days), while a retweet of a tweet in the pre-pandemic dataset had to occur by March 15th. To fall into the post-declaration dataset, a tweet had to occur after March 11th and before May 22nd (74 days), while a retweet had to occur by May 26th.

To quantify the impression of closeness and distance of several measures, we used hierarchical clustering on a cosine distance matrix, using the scipy \cite{Virtanen2019} implementation of the UPGMA algorithm.

For the network graph, the input vectors were retweets for each node in the graph, weighted by the number of retweets, one vector per cluster. For the overall linguistic analyses, vectors were a tf-idf representation of the cleaned corpus, with the collected tweets of each group as documents and a minimum document frequency of 0.5.

For the LIWC analysis for ‘specific content dictionaries’ we used the following columns: ['Posemo', 'Anx', 'Anger', 'Sad', ...]
ADDITIONAL LIWC ANALYSES

B.1 Raw changes

Figure 11 shows the changes in raw score corresponding to the % of max change scores in Figure 8 of the main text.

**FIGURE 11** Raw changes in LIWC Specific Dictionary
C.1 | Gini Coefficients

Per the main paper, we calculated the Gini coefficient - pre- and post-declaration - for the five communities, for a total of ten results. We used a 'weighted' version of the Gini coefficient, that is to say that multiple retweets by y (of a given user x's posts) are counted individually to provide additional 'weight' to the influence of x. This technique is not new: prior work in applying the Gini coefficient on structures of social networks include scientific collaboration networks [Lopes et al., 2011] to health networks [van Mierlo et al., 2016].

Table 3 illustrates our results. The Gini coefficients of all communities, when taken as a whole (across both time periods), are consistently less than the post-declaration coefficients.

Table 3  Changes in absolute numbers of retweets, expressed in 1000s of tweets, with ratio in parentheses.

<table>
<thead>
<tr>
<th>Community</th>
<th>Pre-declaration only</th>
<th>Post-declaration only</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.9791</td>
<td>0.9285</td>
</tr>
<tr>
<td>Democrats</td>
<td>0.9728</td>
<td>0.9364</td>
</tr>
<tr>
<td>Republicans</td>
<td>0.9855</td>
<td>0.9389</td>
</tr>
<tr>
<td>Unorthodox</td>
<td>0.9805</td>
<td>0.8860</td>
</tr>
<tr>
<td>Public Health</td>
<td>0.9490</td>
<td>0.8706</td>
</tr>
<tr>
<td>Antivaxxers</td>
<td>0.9796</td>
<td>0.9450</td>
</tr>
</tbody>
</table>

Note that the result is robust to the weighted/unweighted decision point: even if the unweighted version was used, the Gini coefficients of all communities will still drop in value across time slices.

Pre-pandemic declaration, the highest Gini coefficients were observed in the Republicans and Unorthodox communities. Post-declaration, all communities exhibited decreased centralization, with Republicans and Antivaxxers now exhibiting the highest Gini coefficients. The lowest coefficient both pre- and post-pandemic declaration was in the Health community, suggesting that members of this community distribute their attention more equitably.