Information Processing and Management 59 (2022) 102804

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Investigating Facebook's interventions against accounts that repeatedly share misinformation

Héloïse Théro *, Emmanuel M. Vincent

médialab - Sciences Po, Paris, France

ARTICLE INFO

Keywords:
Misinformation
Content moderation
Algorithmic transparency
Facebook
Fact-checking
Social media analysis

ABSTRACT

Like many web platforms, Facebook is under pressure to regulate misinformation. According to the company, users that repeatedly share misinformation ('repeat offenders') will have their distribution reduced, but little is known about the implementation or the impacts of this measure. The first contribution of this paper is to offer a methodology to investigate the implementation and consequences of this measure, which relies on an analysis combining fact-checking and engagement metrics data. Using a Science Feedback and a Social Science One (Condor) datasets, we identified a set of public accounts (groups and pages) that have shared misinformation repeatedly during the 2019–2020 period. We find that the engagement per post decreased significantly for Facebook pages after they shared two or more 'false news'. The median decrease for pages identified with the Science Feedback dataset is −43%, while this value reaches −62% for pages identified using the Condor dataset. In a different approach, we identified a set of pages claiming to be under 'reduced distribution' for repeatedly sharing misinformation and having received a notification from Facebook. With this set of pages, we observed a median decrease of −25% in engagement per post averaged over 30 days after receiving the notification minus 30 days before. We show that this 'repeat offenders' penalty did not apply to Facebook groups. Instead, we discover that groups have been affected in a different way with a sudden drop in their average engagement per post that occurred around June 9, 2020. While this drop has cut the groups' engagement per post in about half, this decrease was compensated by the fact that these accounts have doubled their number of posts between early 2019 and summer 2020. The net result is that the total engagement on posts from 'repeat offender' accounts (including both pages and groups) returned to its early 2019 levels. Overall, Facebook's policy thus appears to be able to contain the increase in misinformation shared by 'repeat offenders' rather than to decrease it.

1. Introduction

The general public is increasingly getting news related information online, through search engines, social media and video platforms (Mitchell, Gottfried, Barthel, & Shearer, 2016). Hence the spread of misinformation through these platforms has recently received growing attention. Recent studies, along with the political context of January 2021 in the United States, show how the presence of misinformation online can contribute to negative societal consequences. Namely it can fuel false beliefs, such as the idea of a massive voter fraud during the US 2020 presidential election, which may have led to the January 6, 2021 insurrection at the U.S. Capitol (Benkler et al., 2020) and other false stories about presidential candidates (Allcott & Gentzkow, 2017). Misinformation

* Corresponding author.
E-mail addresses: thero.heloise@gmail.com (H. Théro), emmanuel.vincent@sciencespo.fr (E.M. Vincent).

https://doi.org/10.1016/j.ipm.2021.102804
Received 20 January 2021; Received in revised form 26 September 2021; Accepted 23 October 2021
Available online 26 November 2021

has also contributed to confusing the public about the reality of climate change (Brulle, 2018; Porter, Wood, & Bahador, 2019) and stoked skepticism about vaccine safety among the public (Featherstone & Zhang, 2020; Lahouati, De Coucy, Sarlangue, & Cazanave, 2020; Pierri et al., 2021). In April 2020, a questionnaire from the Reuters Institute found that people in the UK use online sources more often than offline sources when looking for information about the coronavirus. Among social media platforms, Facebook was the most widely used with 24% of the respondents saying they used Facebook to access COVID-19 information in the last seven days (Fletcher, Kalogeropoulos, Simon, & Nielsen, 2020). The importance of Facebook in the media landscape is confirmed by Parse.ly's dashboard, which shows that 25% of the visitors of 2500+ media websites are referred by Facebook (Parse.ly's Network Referrer Dashboard, 2021).

Lawmakers and regulators are increasingly pressuring platforms to limit the spread of misinformation. In the US, the House of Representatives organized hearings and convened representatives of the main platforms to testify on how they are being weaponized to spread "misinformation and conspiracy theories online" (Donovan, Jankowicz, Otis, & Smith, 2020). In Europe, the European Commission has established a 'Code of Practice on Disinformation' (Code of practice on disinformation, 2021) that enjoins platforms to voluntarily comply with a set of commitments (Heldt, 2019). Platforms' compliance with the Code of Practice is subjected to an annual assessment by the Commission, the first of which was released in September 2020 (Assessment of the Code of Practice on Disinformation - achievements and areas for further improvement, 2020). The actions that platforms claim to be taking include limiting political advertisement or providing transparency regarding who is funding political advertising, promoting 'authoritative' sources of information, providing data for researchers, sponsoring media literacy initiatives or informing users when they are interacting with misinformation (Annual self-assessment reports of signatories to the Code of Practice on Disinformation, 2019). However, there is little data available and few established processes to monitor the implementation of these measures and quantify their actual impact. Here we propose a methodology to monitor Facebook's implementation of its policy to reduce the visibility of accounts repeatedly spreading misinformation. We chose to focus on Facebook as it is the biggest social media platform with more than two billion users worldwide.

Facebook announced a three-part policy to address 'misleading or harmful content': they claim to *remove* harmful information, *reduce* the spread of misinformation and *inform* people with additional context (Lyons, 2018). Facebook has developed the most extensive third-party fact-checking program with dozens of partner institutions to assist the company in this endeavour (Rosen, 2020). Fact-checkers have access to a stream of viral and likely problematic content, which they can verify and flag as misinformation, with options ranging from "True" (not misinformation), to "Missing context" to "Partly false" and "False" (Rating options for fact-checkers, 2021). Facebook informs page or group owners when published posts on their pages or groups are marked as misinformation, inviting them to correct the posts. The platform's users receive a notification when they have shared a post marked as misinformation and see a notice linking to the fact-check over the flagged posts. A handful of papers provide evidence that supports the efficacy of fact-checking labels by reducing the likelihood that users share false information (Mena, 2020) and reducing false beliefs (Porter & Wood, 2021). In an experimental setting, Pennycook et al. (2020) show that prompting people to consider the accuracy of a piece of information increases the quality of the information they subsequently share on social media. Facebook states that the virality of the posts marked as 'False' or 'Partly False' will be reduced.

The *reduce* policy is not only applied to individual posts, but also to organizations that often publish posts containing misinformation, according to statements in Facebook's publishers help center (Facebook's enforcement of fact-checker ratings, 2021; Fact-checking on Facebook, 2021):

> *Pages and websites that repeatedly share misinformation rated False or Altered will have some restrictions, including having their distribution reduced.*

Facebook ranks each post in users' newsfeed by assigning a relevance score to it. A high score leads to a high likelihood of the post appearing at the top of a user's newsfeed. By decreasing the relevance score, Facebook can make a post or an entire account less visible (Lyons, 2018). However, Facebook has not provided data showing how their *reduce* policy is implemented that would allow researchers to quantify its impact on the spread of misinformation.

One study analyzed the reach of a set of websites identified as sources of false stories on Facebook and Twitter from January 2015 to July 2018. They found that during the 2016 American elections, total engagement on Facebook and Twitter for these sites had more than doubled compared to pre-election levels. Following the election, however, Facebook engagements fell sharply, while Twitter shares continued to increase for these sites, suggesting that Facebook might have taken measures to contain misinformation while Twitter did not (Allcott, Gentzkow, & Yu, 2019).

A more recent article by Kornbluh, Weiner, and Goldstein (2020) measured the level of interactions on Facebook with articles from outlets that repeatedly publish false content from 2016 to 2020, and found results contrasting with those of Allcott et al. (2019). Although Kornbluh et al. did observe a decrease in the first and second quarter of 2017, they observed that total interactions with 'deceptive' outlets on Facebook have increased since then, and were 242% higher during the third quarter of 2020 than during the run-up to the 2016 election. These results suggest that Facebook's policy did not make a lasting impact on misinformation.

Another similar approach was developed by Resnick, Ovadya, and Gilchrist (2018) in the form of the Iffy quotient: a daily calculation of the fraction of the 5,000 most popular URLs on a platform that came from 'iffy' sites (made of a large list of sites that are defined as frequent sources of misinformation and hoaxes). According to this quotient, the proportion of top viral links on Facebook from 'iffy' websites was about 20% during both the 2016 and 2020 US presidential elections. On Twitter, the Iffy quotient increased from about 15% in late October 2016 to around 20% in late October 2020 (Iffy Quotient, 2021). The three studies mentioned above find rather different results due to different methodologies and use of sources that they labelled 'unreliable', but they paint a picture that is in agreement with a persistence of misinformation on Facebook and Twitter at an elevated level.

The present research article departs from articles studying the overall levels of misinformation on platforms by focusing on monitoring a specific policy against misinformation. To that end, we used CrowdTangle, a public insights tool owned and operated by Facebook, to access Facebook data (CrowdTangle Team, 2021). CrowdTangle exclusively tracks public content, and provides access to engagement metrics (such as the number of likes, shares and comments), but not to the reach (number of views) of content (Shiffman, 2021). If Facebook decreases the visibility of posts from accounts sharing misinformation, we expect that their reach decreases. As less users see these posts, the engagement per post should also decrease. To investigate the effect of the *reduce* policy, we used the engagement per post as a proxy for the visibility of the 'repeat offender' accounts content.

We first combined data from one of Facebook's fact-checking partners (Science Feedback) identifying URLs sharing misinformation and from CrowdTangle tracking engagement metrics of the Facebook accounts that repeatedly shared such misinformation. We then replicated this methodology using a set of URLs marked as misinformation by more than one fact-checking organization obtained from Social Science One (called the 'Condor' dataset). Finally, we investigated the engagement metrics of a set of Facebook pages claiming to be under reduced distribution.

2. Research questions

- Is Facebook's policy aiming to reduce the distribution of misinformation from repeat offenders enforced and can its implementation be verified using available engagement data?
- If implemented, what is the magnitude of the reduction in engagement metrics and how does it affect Facebook groups and pages?
- What is the overall impact of the policy on the spread of misinformation on Facebook, i.e. does it result in a decrease in engagement integrated for all repeat offender's accounts over time?

3. Investigating the reduce policy on accounts repeatedly sharing misinformation (Science Feedback data)

To investigate a possible impact of Facebook's policy against accounts that repeatedly share misinformation, we first identified such accounts using data from Science Feedback, one of Facebook's third-party fact-checking partners (Vincent, 2019). Science Feedback is a fact-checking organization dedicated to verifying the credibility of science-related claims and articles. The organization tracks the most viral press articles or social media posts, invites scientists with domain expertise to evaluate their credibility and publishes explanatory articles for a general audience. It contributes to maintaining a database of URLs where the claims checked have been published or repeated that is available online at open.feedback.org.

3.1. Methods

We obtained from Science Feedback a list of 4,000+ URLs reviewed by its team. The list was obtained on January 4, 2021 and cover links flagged in 2019 and 2020. We relied on the 2,452 URLs marked as 'False', which we refer to as 'false news links', excluding the URLs marked as 'Partly False', 'Missing Context', 'False headlines' or 'True', as well as the URLs marked as 'False' but 'corrected' by the publisher, because these labels do not contribute to the 'repeat offender' status according to Facebook's guidelines. Sharing a URL flagged as 'Altered' also contributes to the 'repeat offender' status (Facebook's enforcement of fact-checker ratings, 2021; Fact-checking on Facebook, 2021), but we found no such rating in Science Feedback data.

Using the '/links' endpoint from the CrowdTangle API, we collected the public Facebook groups and pages that shared at least one false news link between January 1, 2019 and December 31, 2020. Due to the API limitations (https://github.com/CrowdTangle/API/wiki/Links), if a URL was shared in more than 1000 posts, we collected only the 1000 posts that received the highest number of interactions; this means that we miss some of the accounts that generate the least interactions and our results focus on the most prolific accounts. We focused on the accounts that spread misinformation the most often, choosing a threshold of 24 different false news links shared over the two years period.

The corresponding 307 Facebook accounts (289 Facebook groups and 18 Facebook pages) are referred to as 'repeat offenders accounts'. The list of accounts is available on the paper's GitHub repository: https://github.com/medialab/webclim_ipm/blob/master/data/section_1_sf/list_accounts_sf.csv. The repository also contains the code used to collect and analyze the data, and to plot the figures. All the posts they published between January 1, 2019 and December 31, 2020 were collected using the '/posts' endpoint. We calculated the engagement per post by summing the number of comments, shares and reactions (such as 'like', 'love', 'favorite', 'haha', 'wow', 'sad' and 'angry' reactions) that each post has received.

'Repeat offender' accounts are supposed to have their distribution reduced, according to Facebook's official communication, but the precise rule Facebook uses to classify an account as 'repeat offender' is not specified. However, a Facebook's staff indicated to a journalist (Solon, 2020) that:

> *The company operates on a 'strike' basis, meaning a page can post inaccurate information and receive a one-strike warning before the platform takes action. Two strikes in 90 days places an account into 'repeat offender' status.*

Based on this ‘two strikes in 90 days’ rule and the list of strike dates known by Science Feedback, we inferred periods during which each account must have been under repeat offender status. If a post shares a misinformation link which was previously fact-checked as ‘False’, we used the date of the post as the strike date. However, if an account shares a link, which later gets fact-checked as ‘False’, then the fact-check date was used as the strike date. A repeat offender period is defined as any given time in which an account shared two or more ‘false news links’ over the past 90 days (see Fig. 1 for two examples).

The set of accounts analyzed comprises some groups and pages that generate more engagement than others by several orders of magnitude. Because the underlying distribution of the engagement metrics were not Gaussian, we used non-parametric statistical methods. To compare the engagement metrics between different periods, we calculated the percentage change for each account, and tested their difference against zero with a Wilcoxon test. As it compares the sums of ranks, a Wilcoxon test is less likely than a t-test to spuriously indicate significance because of the presence of outliers (Wilcoxon, 1945).

The confidence intervals around the medians are estimated using a non-parametric approach, called bootstrap. In each range, the engagement metrics for the N accounts are called events. We randomly sample N events with replacement, meaning that one event can be selected more than once. From the selected N events, we calculate the median. By repeating this process 1,000 times, we obtain 1,000 values for the median. The upper and lower limit of each error bar (in Figs. 2, 4, 5, 7 and 9) represents the 5% and 95% percentiles of the 1,000 medians.

3.2. Results

Fig. 1 displays the engagement metrics for one ‘repeat offender’ group named ‘Australian Climate Sceptics Group’ and one ‘repeat offender’ page named ‘A Voice for Choice’. The known strike dates appear as red lines at the bottom and the inferred ‘repeat offender’ periods are shaded in red. The average engagement per post varies throughout the past two years, but does not appear to be related with the shift between ‘repeat offender’ and ‘no strike’ periods for ‘Australian Climate Sceptics Group’. For the ‘A Voice for Choice’ page, we observe a decrease in engagement in 2020, as the page repeatedly shared different False URLs, which would have maintained it under ‘repeat offender’ status throughout 2020. We compared the average engagement metrics between the ‘repeat offender’ and the ‘no strike’ periods, expecting a decrease in engagement during the ‘repeat offender’ periods. The percentage change is +61% for ‘Australian Climate Sceptics Group’, and −58% for ‘A Voice for Choice’.

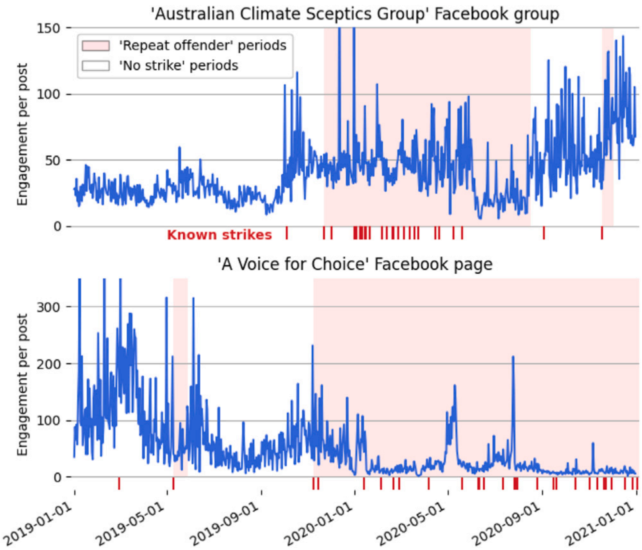

Fig. 1. (Top panel) Average engagement (the sum of comments, shares, reactions) per post for the ‘Australian Climate Sceptics Group’ Facebook group for each day in 2019 and 2020. Each red line at the bottom represents the date of a known strike for this group according to Science Feedback data. The areas shaded in red represent the ‘repeat offender’ periods as defined by the ‘two strikes in 90 days’ rule. **(Bottom panel)** Same as above for the ‘A Voice for Choice’ Facebook page.

To provide a general overview, we calculate the percentage change between the ‘repeat offender’ and the ‘no strike’ periods for each of the 256 Facebook accounts that have published at least one post during each period (see Fig. 2).[1] The median percentage change is −6%, and a Wilcoxon test shows that the values are not significantly different from zero (W = 16051, p-value = 0.74).

When we consider groups and pages separately, the percentage changes are different for the two. For the 238 Facebook groups, the percentage changes are not significantly different from zero (W = 13561, p-value = 0.54), with a median of −3%, while for the 18 Facebook pages, the percentage changes are significantly different from zero (W = 21, p-value = 0.0034), with a median of −43%.

[1] The percentage changes were calculated on the periods between January 1, 2019 and June 8, 2020. Because of the drop in engagement described further down, the second semester of 2020 was excluded (see Fig. 3).

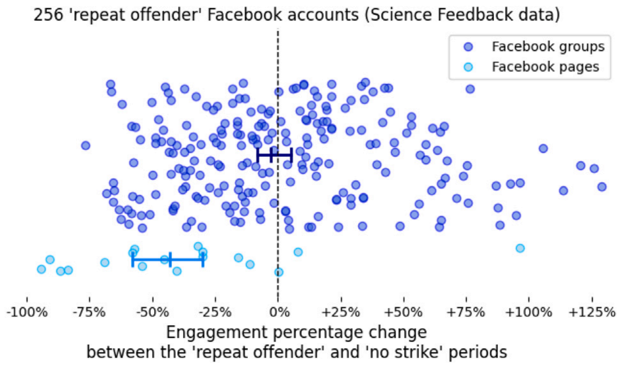

Fig. 2. Percentage changes between the average engagement per post during the 'repeat offender' periods and the 'no strike' periods. Each deep blue dot represents a Facebook group, and each light blue dot a Facebook page. The bars show the medians for each set and their 90% confidence intervals (the intervals are estimated using a bootstrap method). The 256 'repeat offender' accounts represented here were identified using Science Feedback data, and have published at least one post during each period. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

To see whether the strikes would otherwise influence the repeat offenders accounts' engagement over time, we analyzed the total amount of engagement received by all the posts published by each of the 307 repeat offenders accounts for each day of the 2019–2020 period (Fig. 3). This metric, representing the total engagement generated by these accounts on Facebook (top panel), can be decomposed as the number of posts published each day (middle panel) times the average number of engagement per post (bottom panel).

The total engagement per day is stable from January to September 2019, however we observe a rise from September 2019 to June 2020. This rise is explained by the increase in activity of the misinformation accounts (with a doubling of the number of posts per day) while the engagement per post remained rather constant. Around June 9, 2020, the total engagement metrics have massively dropped. This decrease is entirely explained by a corresponding drop in engagement per post (Fig. 3). While this drop has cut the groups' engagement per post in half, this decrease was compensated by the fact that 'repeat offender' accounts have doubled their number of posts between 2019 and 2020. The net result is that the total engagement on posts from 'repeat offender' accounts returned to its early 2019 levels.

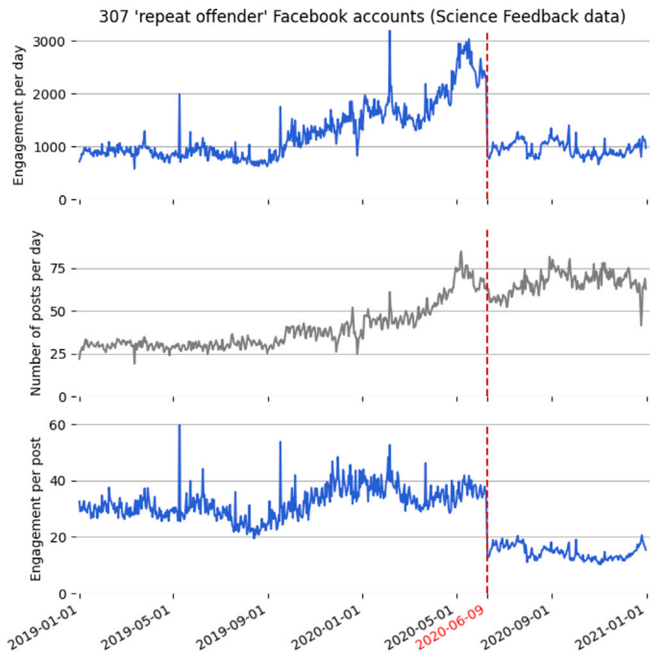

Fig. 3. Metrics aggregated over the 307 Facebook accounts that repeatedly shared false news links identified using Science Feedback data. **(Top panel)** Total engagement per day averaged for all accounts. **(Middle panel)** Number of posts per day averaged for all accounts. **(Bottom panel)** Average engagement per post. The dotted red line marks the date of June 9, 2020, when a sudden drop in engagement is observed.

To further quantify this 'June drop', we calculated the percentage change in engagement per post for each account during a 30-day period before and after June 9, 2020 (Fig. 4). The median percentage change is −43%, and most of the accounts (219 out of 289) experienced a decrease in engagement.[2] A Wilcoxon test indicates that these percentage changes are significantly different from zero ($W = 9012$, p-value $= 4.6 \times 10^{-17}$).

When we consider groups and pages separately, the percentage changes are different for the two. While the percentage changes for the 271 groups are significantly different from zero ($W = 7599$, p-value $= 5.1 \times 10^{-17}$), with a median of −45%, the 18 pages appear to not be affected by the decrease ($W = 73$, p-value $= 0.61$), with a median percentage change of −5%. As the June drop does not affect groups and pages equally, we reproduced Fig. 3's bottom panel for groups and pages separately (see Supplementary Figure S1), which further shows that the June 2020 engagement metrics' drop only affects groups.

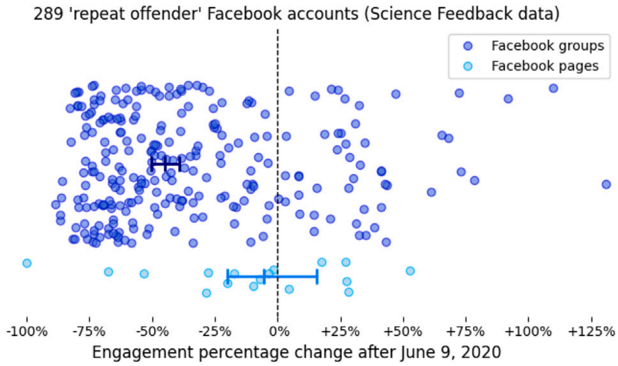

Fig. 4. Percentage changes in the average engagement per post during a 30-day period before and after June 9, 2020. Each deep blue dot represents a Facebook group, and each light blue dot a Facebook page. The bars show the medians for each set and their 90% confidence intervals. The 289 'repeat offender' accounts represented here were identified by Science Feedback data, and have published at least one post one month before and one month after June 9, 2020. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

To verify whether this drop was specific to 'repeat offender' groups, we compared these dynamics to those of a control set of accounts consisting of Facebook pages and groups associated with accounts that did not publish misinformation. No such drop in engagement was observed around June 9, 2020 (see Supplementary Figure S4).

The most likely explanation for such a massive change is that Facebook modified how its algorithm promoted the content from these groups starting on June 9, 2020. While we did observe a relationship between the strike dates and a decrease in engagement for 'repeat offender' pages, we observed no such link for 'repeat offender' groups. Hence it seems that Facebook took action against these groups via this one-shot measure in June 2020.

4. Investigating the reduce policy on accounts repeatedly sharing misinformation (Condor data)

One limitation of the results described above is that we obtained the links labelled as 'False' from only one fact-checking organization (Science Feedback), while Facebook partners with over 80 fact-checking organizations (Rosen, 2020). The accounts might have received strikes from other fact-checkers apart from Science Feedback, which would create longer or additional 'repeat offender' periods. The true 'repeat offender' periods could thus be different than the ones inferred, potentially changing the magnitude of the 'reduce' effect. In this section, we clear up this potential issue by replicating the above analysis using a dataset containing flags from other fact-checking organizations.

4.1. Methods

We use data from the Social Science One organization (King & Persily, 2020), a consortium of research centers that builds partnerships between academia and private companies such as Facebook to share data and expertise. In June 2021, Social Science One released a new version of the Condor dataset (Messing et al., 2021), which contains all the URLs shared on the platform by at least 100 Facebook users between January 1, 2017 and February 28, 2021, as well as their fact-checking metadata. From this list, we extracted the 6,811 URLs that were shared in 2019 or 2020, that were flagged as 'False', and whose country in which it was shared the most was either the USA, Canada, Great Britain or Australia.

We then replicated the data processing described in the previous section. Using CrowdTangle, we collected all the posts that shared one of the false links and identified 706 Facebook accounts (671 Facebook groups and 35 Facebook pages) that shared at least 24 false links between January 1, 2019 and December 31, 2020. Then we used CrowdTangle to collect all the posts published by those accounts in 2019 and 2020. The Condor dataset contains the date of the first fact-check article used to flag a URL. We could thus infer the 'repeat offender' periods for each account and conduct the same analysis as in the previous section.

[2] A decrease in engagement on June 9, 2020 can be seen for the 'Australian Climate Sceptics Group' in Fig. 1 (the percentage change was −60% for this example).

Science Feedback being a third-party fact-checker working with Facebook, some of their URLs are also contained in the Condor dataset (see Supplementary Figure S5). Thus a significant part of the 'repeat offender' groups and pages obtained from the Condor URLs are the same as the ones analyzed previously. As the point of this section's analysis is to test the replicability of the previous results, we chose to exclude the accounts that were analyzed in the previous section. This section hence presents the engagement metrics only for the 503 'novel' accounts, which represent 476 groups and 27 pages (see Supplementary Figure S6). The list of accounts is available here: https://github.com/medialab/webclim_ipm/blob/master/data/section_2_condor/list_accounts_condor.csv.

4.2. Results

Our first objective is to verify whether the repeat offender policy was applied differently to Facebook pages and groups as observed in the previous section. To that end, we calculate the percentage change in engagement between the 'repeat offender' and the 'no strike' periods for each of the 437 Facebook accounts that have published at least one post during each period (see Fig. 5). The median percentage change is −5%, and the values are not significantly different from zero (W = 46495, p-value = 0.61).

As in the previous section, changes in engagement per post are different for groups and pages (Fig. 5). The percentage change for the 414 Facebook groups are not statistically different from zero (W = 41561, p-value = 0.57), with a median of −2%, while the values for the 23 Facebook pages are significantly different from zero (W = 29, p-value = 0.00041), with a median of −62%.

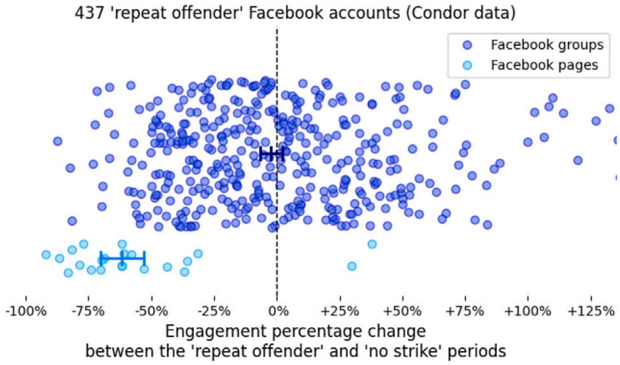

Fig. 5. Same metric as on Fig. 2. The 437 'repeat offender' accounts presented here were identified using the Condor dataset, and have published at least one post during each period.

We then analyzed timeseries of the engagement received by the 503 'repeat offender' accounts in 2019 and 2020 (see Fig. 6). The 'novel' accounts similarly display a gradual rise in total engagement from September 2019 to June 2020, and a massive drop around June 9, 2020. Again, we observe that this 'June drop' resulted in a decrease in engagement per post by about half, and brought the total engagement on posts from 'repeat offenders' back to its early 2019 levels.

To quantify the 'June drop', we calculated the percentage change in engagement for each account during a 30-day period before and after June 9, 2020 (Fig. 7). The median percentage change is −26%, and 63% of the accounts experienced a decrease in engagement. This decrease is smaller than the one observed previously (−43%). The values are still significantly different from zero (W = 42651, p-value = 3.8×10^{-5}).

When we consider groups and pages separately, the percentage changes for the 442 groups are significantly different from zero (W = 37889, p-value = 3.8×10^{-5}) and the median is −27%, whereas the values for the 23 pages are not different from zero (W = 133, p-value = 0.89), with a median of −2% (Fig. 7). When the engagement per post is plotted separately for groups and pages, we again observe a drop in engagement for groups only (see Supplementary Figure S2).

This analysis using a dataset of 'False' URLs flagged by several fact-checking organizations confirms the main findings of the previous section. Pages undergo a period of decrease in engagement following the publication of two false links, while groups have been affected by a sudden decrease in engagement for all their posts in June 2020.

A limitation of these results is that we relied on the strike dates to infer the 'repeat offender' periods, but we cannot know for certain whether the pages investigated were actually under a 'repeat offender' status. Indeed, one could imagine that the 'two strikes in less than 90 days' rule has been modified, or that links fact-checked as 'partly false' or 'missing context' might also be taken into account to determine the repeat offender status. In the next section, we used a different methodology to collect pages for which we are sure that they are under 'repeat offender' status.

5. Investigating the reduce policy on pages declaring to be under 'reduced distribution'

5.1. Methods

We noticed that two popular pages ('Mark Levin' and '100 Percent FED Up') have publicly shared a message claiming to be placed under 'repeat offender' status with a screenshot as a piece of evidence. To gather a list of such self-declared repeat offenders, we searched on CrowdTangle for posts published since January 1, 2020 with the following keywords:

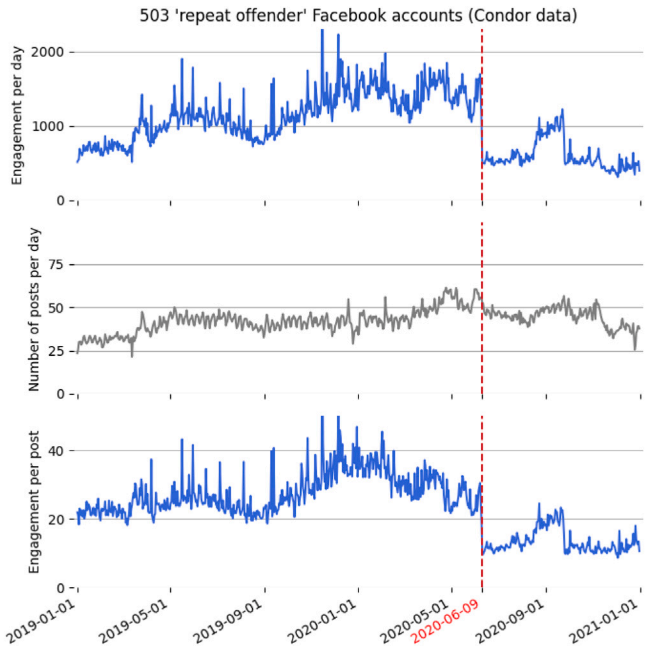

Fig. 6. Same metrics as on Fig. 3 aggregated over the 503 'repeat offender' Facebook accounts identified using Condor data.

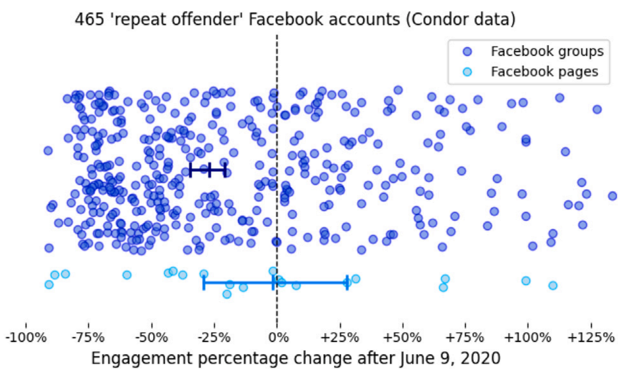

Fig. 7. Same metric as on Fig. 4. The 465 'repeat offender' accounts represented here were identified using the Condor dataset, and have published at least one post one month before and one month after June 9, 2020.

- 'reduced distribution' AND ('restricted' OR 'censored' OR 'silenced')
- 'Your page has reduced distribution'

For this we used the '/posts/search' endpoint of the API on November 25, 2020.

We manually opened the resulting posts, and kept the ones which met the following criteria (see Fig. 8 top panel for an example):

- The post should include a screenshot of the Facebook notification.
- In the screenshot, the Facebook notification should say: 'Your page has reduced distribution and other restrictions because of repeatedly sharing of false news.'
- In the screenshot, the name of the page should be visible.

Doing so, we obtained a list of 94 pages. We found only Facebook pages in this case, and no groups. A search using the terms 'Your group has reduced distribution' did not yield any result.

To verify whether Facebook applied any restriction to these pages, we collected all the posts that these 94 pages have published between January 1, 2019 and December 31, 2020 from the CrowdTangle API using the '/posts' endpoint. The collection was run on January 11, 2021. We were only able to collect data from 83 of these pages, as 11 were deleted from the CrowdTangle database since our search in November 2020. This highlights an important issue when studying misinformation trends on Facebook: some data disappears as accounts are deleted or changed to 'private'.

Among the 83 Facebook pages collected, two were already among the 18 pages included in the first analysis, and one was already present in the set of 35 pages included in the second analysis (see Supplementary Figure S6). We excluded these pages to present only the 80 'novel' pages in this section. The list of accounts is available here: https://github.com/medialab/webclim_ipm/blob/master/data/section_3_self_declared/list_accounts_reduce.csv.

The date of the last fact-check notification was used as the inferred start date of reduced distribution, when it appeared in the screenshot. When it was not visible, we used the date of the post as the inferred start date of reduced distribution. The inferred 'reduced distribution' dates range from April 1rst to November 23, 2020. We are aware that the inferred date may not correspond to the real date at which the restrictions has begun to be enforced. For example, a page may have received a 'reduced distribution' notification from Facebook in early 2020, while sharing a screenshot of this notification only a few days or weeks later. Because the 'reduced distribution' notification is a private message that cannot be accessed unless the page shares it publicly, we had no choice but to rely on this inferred date as a proxy for the start date of the restrictions.

5.2. Results

Fig. 8 shows a screenshot of the Facebook notification shared by the '100 Percent FED Up' page (with a last violation notified on July 31, 2020), and the average engagement per post for that page over the past two years. We observe substantial variations in the metric, as well as a drop in engagement in August 2020. When we compare the average engagement per post during a 30-day period after and before July 31, 2020, the percentage change is −62%.

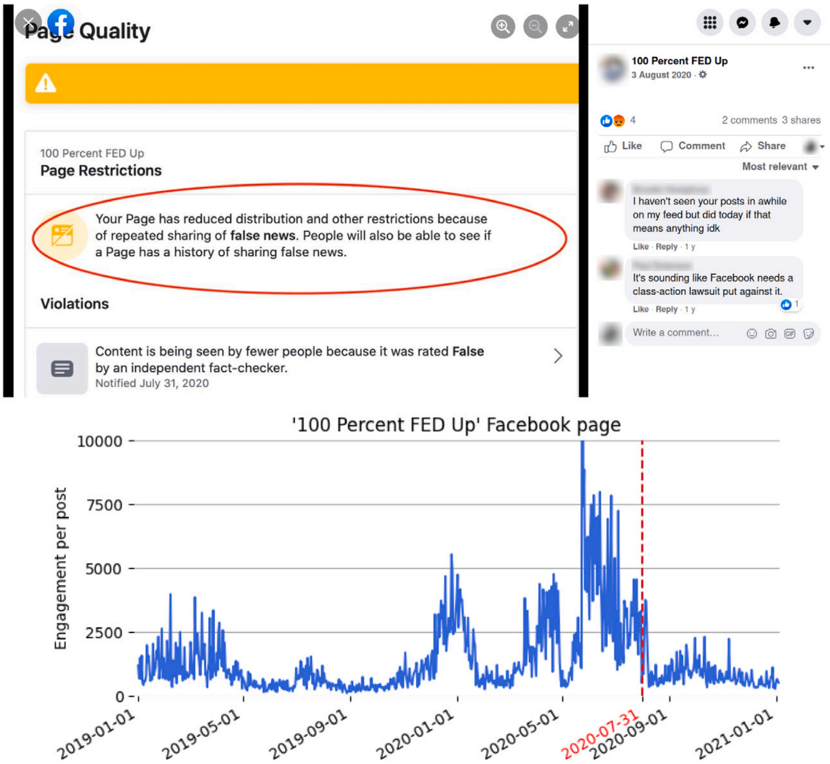

Fig. 8. **(Top panel)** Screenshot of a post from the '100 Percent FED Up' Facebook page sharing a 'reduced distribution' notification from Facebook (screenshot taken on September 22, 2021). **(Bottom panel)** Average engagement per post for the '100 Percent FED Up' page for each day in 2019 and 2020. The dotted red line represents the reduced distribution start date that is inferred from the date of the last violation on the screenshot ('Notified July 31, 2020').

To provide a general overview, we calculate the percentage change in engagement during a 30-day period before and after the reduced distribution start date for each of the 79 Facebook pages that published at least one post during each period (see Fig. 9). The median percentage change is −25%, and a Wilcoxon test reveals that the percentage changes are significantly different from zero (W = 855, p-value = 0.00040). The 'reduced distribution' notification does appear to be followed by a modest decrease in engagement per post.

Finally, we verify whether an important drop in engagement also occurred in June 2020 for this set of Facebook pages. When we compare the engagement metrics before and after June 9, 2020, the median percentage change is 4%. Although the difference from zero is marginally significant (W = 992, p-value = 0.049), it means that the engagement per posts tended to *increase* after June 2020 for these pages (see Supplementary Figure S3). This further confirms that Facebook pages have not been affected by the sudden *reduce* measure implemented in June 2020 as evidenced in the previous sections.

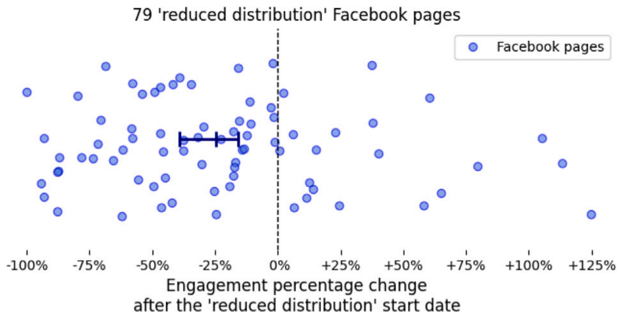

Fig. 9. Percentage changes in average engagement per post during a 30-day period after minus before the reduced distribution notification date. Each dot represents a Facebook page. The bars show the median and its 90% confidence interval. The 79 'reduced distribution' pages presented here were identified because they shared a 'reduced distribution' notification from Facebook.

6. Discussion

Facebook, the most widely used social media platform in the world, has announced a series of measures to curb the spread of misinformation, notably by reducing the visibility of content shared by 'repeat offenders', which are accounts that repeatedly share false information. However, the effects of the platforms' diverse policies to tackle misinformation remains understudied (Pasquetto et al., 2020). The present research article aims to contribute to filling this knowledge gap by providing a method to verify the application and measure the consequences of Facebook's 'reduce' policy on the targeted accounts' engagement metrics.

As a first step, we investigated 307 Facebook accounts (mainly groups) having repeatedly shared misinformation using a fact-checker's dataset. Sharing two false links over a three-month period is supposed to be penalized by a reduced visibility of the account's content (Solon, 2020). We did observe a significant decrease (median of −43%) in the engagement per posts published by pages under a presumptive repeat offender status. However, we find that this policy is not leading to a significant decrease in engagement for Facebook groups.

As a second step, we replicated this methodology using another dataset of URLs shared by Facebook, and identified 503 additional accounts that have shared misinformation repeatedly. We again observed a significant decrease (median of −62%) in engagement per post for 'repeat offender' pages, while the engagement for 'repeat offender' groups was not affected.

As a third step, we identified 83 Facebook pages which have shared a Facebook notification, indicating that their account was under reduced distribution. The pages' engagement metrics were significantly lower after the date of the notification (median of −25%), suggesting that the 'reduced distribution' measure was indeed applied to the pages after they received a notification from Facebook. We noted that no group was found when searching for accounts sharing a reduced distribution notification, which confirms that the 'repeat offender' policy is applied only to Facebook pages, and not to groups.

The different methodologies we used to infer the repeat offender periods are subjected to biases. First, to perfectly identify the repeat offender periods for a given account, we should have knowledge of all the False URLs this account has shared. The two URL datasets we relied on have different limitations in this regard: the Condor dataset contains URLs from all the fact-checkers Facebook is working with, but only the URLs shared more than 100 times on its platform, while the Science Feedback dataset contains all URLs they flagged, regardless of sharing numbers, but only for this one fact-checking organization. Moreover the 'two strikes in 90 days' rule used to infer repeat offender periods may not be the one used by Facebook as of today, and it may not be the only one. Second, the time when a page shared their 'reduced distribution' notification may be days or weeks after the page actually received the notification.

Given that there is no public data indicating when a group or page is under repeat offender status, the only way to monitor the effects of the repeat offender's policy was to infer which groups or pages should be under this status from the data available to us. The three approaches described above yield consistent results for a decrease in engagement per post for repeat offender pages, with a median decrease in engagement ranging from −62% to −25%, suggesting that we were able to capture, at least partially, the repeat offender periods. There might exist other potential biases, such as a few days lag before enforcement of the policy to allow for human verification. In any case, such issues might only result in underestimating the true size effect of the 'repeat offender' interventions by not contrasting exactly the real 'repeat offender' and 'no strike' periods.

Although we observe an overall reduction in engagement, there exists a large heterogeneity across the 'repeat offender' pages (see Figs. 2, 5 and 9). The engagement per post has actually increased for some popular pages, such as the 'Tucker Carlson Tonight' page with a 38% increase (from 104k to 143k interactions per post) following the 'reduced distribution' notification from Facebook, for example. The engagement per post for the 'Mark Levin' page remained rather stable following the notification, changing from 20.3k to 20.7k interactions per post (a 2% change). It is possible that these accounts compensated for the decrease in engagement due to the 'reduce' intervention with a simultaneous gain in followers or an increase in their motivation for engaging with their content. Another possible explanation could be that these high-profile users are not affected by Facebook's policies in the same way as others. A recent investigation published by the Wall Street Journal suggests the existence of distinct enforcement procedures for high-profile Facebook users, such as celebrities, politicians and journalists (Horwitz, 2021). At this stage, this is only a hypothesis that would deserve further investigations.

By analyzing the time series of the repeat offenders' engagement over the past two years, we also discovered a sudden drop in engagement affecting the groups around June 9, 2020. For many groups, the decrease was quite drastic (up to −70% - −80%), with a median drop in engagement of −45% for the first analysis and −27% for the second one. The 18 Facebook pages from the first sample, the 23 pages from the second sample, as well as the 80 pages from the third sample, were not affected by this decrease. This 'June drop' does not correspond to any official communication by Facebook on that matter. It indicates that the company has very likely taken internal decisions that heavily impact the organic reach of most repeat offender groups, in ways that are not stated in public announcement of the company's policies. More transparency from Facebook would be needed to understand the nature and origin of this change. It would also bring clarity on how rules aimed at limiting the spread of misinformation are being enforced.

It is not clear why only repeat offender Facebook groups, and not pages, saw their engagement reduced in June 2020. Studies have highlighted that misinformation persists at high levels on Facebook and other platforms (Kornbluh et al., 2020; Resnick et al., 2018). In the context of the COVID-19 pandemic, concerns rose about the amount of misinformation spreading on social media, including Facebook, and its potential harm to users (Johnson et al., 2020). It is possible that such concerns have driven Facebook to apply a 'quick fix' to decrease the engagement of posts shared in groups spreading misinformation and compensate for the absence of a repeat offender policy. One should note that since the overall activity in these misinformation groups doubled between September 2019 and June 2020, the 'June drop' has only succeeded in bringing the overall engagement level back to its early 2019 values (see Figs. 3 and 6 top panels).

Facebook pages and groups have different purposes: pages are meant to be for official communication from the page administrators to a large audience, while groups are meant to foster interactions between users (What's the difference between a profile, page and group on Facebook?, 2021). Pages are thus always public, while groups can be public or private. Pages' posts can also be monetized and promoted. Despite these differences, we have seen that both pages and groups are being used to share false news, and we actually found vastly more groups than pages when we identified the accounts spreading the most misinformation. Indeed, groups represented 94% of the accounts sharing at least 24 False URLs in the first analysis, and 95% in the second analysis. In the interest of curbing the spread of misinformation, applying its 'repeat offender' policy to groups as well as to pages would have helped Facebook to decrease the amount of misinformation in their users' feeds in 2019 and 2020.

It should be noted that platforms only started fighting against misinformation recently, following the 2016 U.S. presidential election for some and the onset of the COVID-19 pandemic for others, and their practices are still evolving. We note that Facebook appears to have started applying its 'repeat offenders' policy on misinformation groups in the spring of 2021. Indeed, one of the 'repeat offender' groups analyzed here has shared a 'reduced distribution' notification from Facebook in May 2021 (see Fig. 10). Furthermore, Facebook announced in May 2021 that individuals' Facebook accounts will now also have reduced distribution if they repeatedly share misinformation (Taking action against people who repeatedly share misinformation, 2021). It would thus be interesting to replicate our approach on engagement data for 2021 to monitor the effects of these new measures.

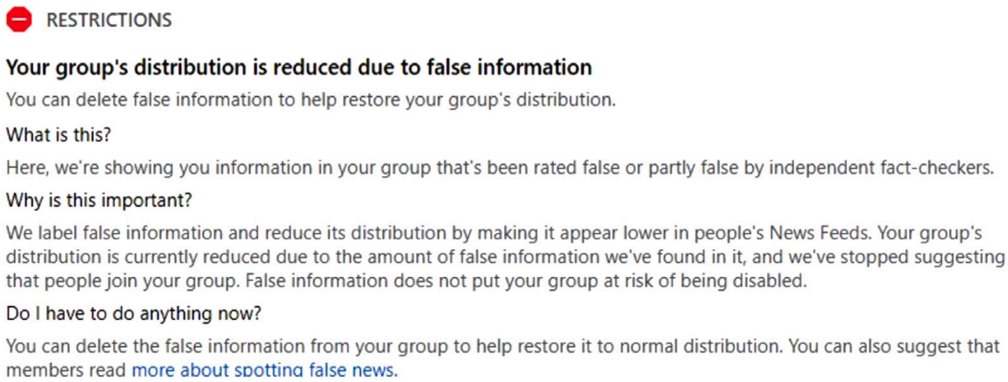

Fig. 10. Screenshot of the post of a 'repeat offender' group, sharing in May 2021 a 'reduced distribution' notification sent by Facebook.

It remains possible for both groups and pages to evade the consequences of sharing false information repeatedly. Groups are informed by Facebook that they "can delete false information to help restore [their] group's distribution" (see Fig. 10), which makes it straightforward to escape the repeat offender status. By contrast, page owners cannot get rid of a strike as easily. Facebook informs them that: "deleting a post will not eliminate the strike against the Page or domain". Instead, they have to correct the offending posts and submit an appeal to the fact-checker for the strike to be lifted (Issue a correction or dispute a rating, 2021). All members of a group, and not just its administrators, can publish posts, making it hard for owners to control it editorially. That may be why group owners can restore their status by deleting a post flagged by fact-checkers, and why groups were not affected by the repeat offender policy until 2021.

Online misinformation can be a threat to society, and the role that platforms can play via targeted interventions, has been the subject of intense debate over the past few years (Rogers, 2020). As a consequence, researchers (Mena, 2020; Yaqub, Kakhidze, Brockman, Memon, & Patil, 2020) and journalists (Facebook offers a distorted view of American news, 2020; Roose, Isaac, & Frenkel, 2020) have begun to monitor the actions that platforms take to tackle misinformation and their efficacy. Given the facts that:

(1) false news go viral much faster than fact-checks can get published,
(2) accounts that have shared misinformation in the past tend to keep sharing misinformation,

(3) a small number of accounts is responsible for a large proportion of the misinformation being shared (at least regarding COVID-19 (The disinformation dozen: Why platforms must act on twelve leading online anti-vaxxers, 2021)),

then acting against 'repeat offenders' is likely to be one of the most effective interventions that platforms can make to protect their users against manipulation.

There is a critical need for further research to thoroughly verify and shed light on platforms' actions against misinformation. While our results provide information on the relative drop in engagement per post resulting from Facebook's repeat offenders policy, more research is needed to quantify the impact of such policies on the overall prevalence of misinformation in users' feeds.

CRediT authorship contribution statement

Héloïse Théro: Data curation, Investigation, Software, Validation, Visualization, Writing – original draft. **Emmanuel M. Vincent:** Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – review & editing.

Acknowledgments

We are very grateful to Social Science One and to Facebook for their partnership in making the Condor Facebook URL Shares dataset available to researchers including ourselves. We thank Shaden Shabayek, Manon Berriche and two anonymous reviewers for their insights on the manuscript. We also thank Guillaume Plique, Benjamin Ooghe-Tabanou and all the médialab technical team for their help with data collection. This research was supported by the 'Make Our Planet Great Again' French state aid managed by the Agence Nationale de la Recherche under the 'Investissements d'avenir' program with the reference ANR-19-MPGA-0005.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ipm.2021.102804.

References

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236. http://dx.doi.org/10.1257/jep.31.2.211.

Allcott, H., Gentzkow, M., & Yu, C. (2019). Trends in the diffusion of misinformation on social media. *Research & Politics*, *6*(2), Article 2053168019848554. http://dx.doi.org/10.1177/2053168019848554.

Annual self-assessment reports of signatories to the Code of Practice on Disinformation 2019 (2019). European Commission. https://digital-strategy.ec.europa.eu/en/news/annual-self-assessment-reports-signatories-code-practice-disinformation-2019.

Assessment of the Code of Practice on Disinformation - achievements and areas for further improvement (2020). European Commission. https://digital-strategy.ec.europa.eu/en/library/assessment-code-practice-disinformation-achievements-and-areas-further-improvement.

Benkler, Y., Tilton, C., Etling, B., Roberts, H., Clark, J., Faris, R., et al. (2020). Mail-in voter fraud: Anatomy of a disinformation campaign. The Berkman Klein Center for Internet & Society at Harvard University. https://cyber.harvard.edu/publication/2020/Mail-in-Voter-Fraud-Disinformation-2020.

Brulle, R. (2018). 30 Years ago global warming became front-page news–and both Republicans and democrats took it seriously. The Conversation. https://theconversation.com/30-years-ago-global-warming-became-front-page-news-and-both-republicans-and-democrats-took-it-seriously-97658.

Code of Practice on Disinformation (2021). European Commission. https://ec.europa.eu/digital-single-market/en/code-practice-disinformation.

CrowdTangle Team (2021). CrowdTangle. Facebook, Menlo Park, California, United States. List ID: 1421627, 1422062, 1466638, 1480255, 1491244, 1491266, 1491267, 1491268, 1492390, 1491269, 1590764, 1591619, 1592120, 1592111, 1593557, 1593558.

Donovan, J., Jankowicz, N., Otis, C., & Smith, M. (2020). House Intelligence Committee open virtual hearing: "Misinformation, conspiracy theories, and 'infodemics': Stopping the spread online". https://intelligence.house.gov/news/documentsingle.aspx?DocumentID=1092.

Facebook offers a distorted view of American news (2020). The Economist. https://www.economist.com/graphic-detail/2020/09/10/facebook-offers-a-distorted-view-of-american-news.

Facebook's enforcement of fact-checker ratings (2021). Facebook Help. https://www.facebook.com/business/help/297022994952764.

Fact-checking on Facebook (2021). Facebook Help. https://www.facebook.com/business/help/2593586717571940.

Featherstone, J. D., & Zhang, J. (2020). Feeling angry: the effects of vaccine misinformation and refutational messages on negative emotions and vaccination attitude. *Journal of Health Communication*, *25*(9), 692–702. http://dx.doi.org/10.1080/10810730.2020.1838671.

Fletcher, R., Kalogeropoulos, A., Simon, F. M., & Nielsen, R. K. (2020). Information inequality in the UK coronavirus communications crisis. Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/information-inequality-uk-coronavirus-communications-crisis.

Heldt, A. (2019). Let's meet halfway: Sharing new responsibilities in a digital age. *Journal of Information Policy*, *9*, 336–369. http://dx.doi.org/10.5325/jinfopoli.9.2019.0336.

Horwitz, J. (2021). Facebook says its rules apply to all. Company documents reveal a secret elite that's exempt. The Wall Street Journal. https://www.wsj.com/articles/facebook-files-xcheck-zuckerberg-elite-rules-11631541353.

Iffy quotient (2021). https://csmr.umich.edu/projects/iffy-quotient/.

Issue a correction or dispute a rating (2021). Facebook Help. https://www.facebook.com/business/help/997484867366026.

Johnson, N. F., Velásquez, N., Restrepo, N. J., Leahy, R., Gabriel, N., El Oud, S., et al. (2020). The online competition between pro-and anti-vaccination views. *Nature*, *582*(7811), 230–233. http://dx.doi.org/10.1038/s41586-020-2281-1.

King, G., & Persily, N. (2020). A new model for industry–academic partnerships. *PS: Political Science & Politics*, *53*(4), 703–709. http://dx.doi.org/10.1017/S1049096519001021.

Kornbluh, K., Weiner, E., & Goldstein, A. (2020). New study by digital new deal finds engagement with deceptive outlets higher on Facebook today than run-up to 2016 election. GMF The German Marshall Fund of the United States. https://www.gmfus.org/news/new-study-digital-new-deal-finds-engagement-deceptive-outlets-higher-facebook-today-run-2016.

Lahouati, M., De Coucy, A., Sarlangue, J., & Cazanave, C. (2020). Spread of vaccine hesitancy in France: What about YouTube™? *Vaccine*, *38*(36), 5779–5782. http://dx.doi.org/10.1016/j.vaccine.2020.07.002.

Lyons, T. (2018). The three-part recipe for cleaning up your news feed. Facebook Newsroom. https://about.fb.com/news/2018/05/inside-feed-reduce-remove-inform/.

Mena, P. (2020). Cleaning up social media: The effect of warning labels on likelihood of sharing false news on Facebook. *Policy & Internet*, *12*(2), 165–183. http://dx.doi.org/10.1002/poi3.214.

Messing, S., DeGregorio, C., Hillenbrand, B., King, G., Mahanti, S., Mukerjee, Z., et al. (2021). *Facebook privacy-protected full urls data set*. Havard Dataverse, V7. http://dx.doi.org/10.7910/DVN/TDOAPG.

Mitchell, A., Gottfried, J., Barthel, M., & Shearer, E. (2016). The modern news consumer: News attitudes and practices in the digital era. Pew Research Center. https://www.pewresearch.org/journalism/2016/07/07/the-modern-news-consumer/.

Parse.ly's Network Referrer Dashboard (2021). https://www.parse.ly/resources/data-studies/referrer-dashboard. (Accessed 8 July 2021).

Pasquetto, I. V., Swire-Thompson, B., Amazeen, M. A., Benevenuto, F., Brashier, N. M., Bond, R. M., et al. (2020). Tackling misinformation: What researchers could do with social media data. *Harvard Kennedy School (HKS) Misinformation Review*, *1*(8), http://dx.doi.org/10.37016/mr-2020-49.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A., Eckles, D., & Rand, D. (2020). Understanding and reducing the spread of misinformation online. *Advances in Consumer Research*, *48*, 863–867.

Pierri, F., Perry, B., DeVerna, M. R., Yang, K.-C., Flammini, A., Menczer, F., et al. (2021). The impact of online misinformation on US COVID-19 vaccinations. arXiv preprint arXiv:2104.10635.

Porter, E., & Wood, T. J. (2021). The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom. *Proceedings of the National Academy of Sciences*, *118*(37), http://dx.doi.org/10.1073/pnas.2104235118.

Porter, E., Wood, T. J., & Bahador, B. (2019). Can presidential misinformation on climate change be corrected? evidence from internet and phone experiments. *Research & Politics*, *6*(3), Article 2053168019864784. http://dx.doi.org/10.1177/2053168019864784.

Rating options for fact-checkers (2021). Facebook Help. https://www.facebook.com/business/help/341102040382165.

Resnick, P., Ovadya, A., & Gilchrist, G. (2018). Iffy quotient: A platform health metric for misinformation, vol. 17. Center for Social Media Responsibility. http://umsi.info/iffy-quotient-whitepaper.

Rogers, R. (2020). Deplatforming: Following extreme internet celebrities to telegram and alternative social media. *European Journal of Communication*, *35*(3), 213–229. http://dx.doi.org/10.1177/0267323120922066.

Roose, K., Isaac, M., & Frenkel, S. (2020). Facebook struggles to balance civility and growth. The New York Times. https://www.nytimes.com/2020/11/24/technology/facebook-election-misinformation.html.

Rosen, G. (2020). An update on our work to keep people informed and limit misinformation about COVID-19. Facebook Newsroom. https://about.fb.com/news/2020/04/covid-19-misinfo-update/.

Shiffman, N. (2021). Understanding and citing CrowdTangle data. CrowdTangle Communication. https://help.crowdtangle.com/en/articles/4558716-understanding-and-citing-crowdtangle-data.

Solon, O. (2020). Sensitive to claims of bias, facebook relaxed misinformation rules for conservative pages. NBC News. https://www.nbcnews.com/tech/tech-news/sensitive-claims-bias-facebook-relaxed-misinformation-rules-conservative-pages-n1236182.

Taking action against people who repeatedly share misinformation (2021). Facebook Newsroom. https://about.fb.com/news/2021/05/taking-action-against-people-who-repeatedly-share-misinformation/.

The disinformation dozen: Why platforms must act on twelve leading online anti-vaxxers (2021). Center for Countering Digital Hate. https://www.counterhate.com/disinformationdozen.

Vincent, E. M. (2019). Science Feedback partnering with facebook in fight against misinformation. Science Feedback. https://sciencefeedback.co/science-feedback-partnering-with-facebook-in-fight-against-misinformation/.

What's the difference between a profile, page and group on Facebook? (2021). Facebook Help Centre. https://www.facebook.com/help/337881706729661/.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80–83. http://dx.doi.org/10.2307/3001968.

Yaqub, W., Kakhidze, O., Brockman, M. L., Memon, N., & Patil, S. (2020). Effects of credibility indicators on social media news sharing intent. In *Proceedings of the 2020 Chi conference on human factors in computing systems* (pp. 1–14). http://dx.doi.org/10.1145/3313831.3376213.