Investigating Facebook's interventions against accounts that repeatedly share misinformation

Héloïse Théro[a,*], Emmanuel M. Vincent[a,*]

[a] *médialab - Sciences Po, Paris, France*

Abstract

Like many web platforms, Facebook is under pressure to regulate misinformation. According to the company, users that repeatedly share misinformation ('repeat offender') will have their distribution reduced, but little is known about the implementation or the impacts of this measure. This paper investigates the implementation and consequences of this policy using a first of its kind analysis, combining data from a fact-checking organization (Science Feedback), Facebook's Social Science One dataset (Condor), users' self-declaration and CrowdTangle data. Based on Science Feedback data, we first identified a set of public accounts (groups and pages) that have shared misinformation repeatedly during the 2019-2020 period. The engagement per post decreased for Facebook pages after they shared two or more 'false news', and this result was replicated using Condor data. We also discover that Facebook groups have been affected in a different way with a sudden drop in their average engagement per post that occurred around June 9, 2020. Finally we identified a set of pages claiming to be under 'reduced distribution' by Facebook for repeatedly sharing misinformation, and we again observed a decrease in their engagement per post. In the three sets of pages studied, the median decrease in engagement after sharing misinformation is ranging from -62% to -24%.

Keywords: Misinformation, Content moderation, Algorithmic transparency,

[*] Corresponding authors.

Email addresses: thero.heloise@gmail.com (Héloïse Théro), emmanuel.vincent@sciencespo.fr (Emmanuel M. Vincent)

Facebook, Fact-checking, Social media analysis

1. Introduction

The general public is increasingly getting news related information online, through search engines, social media and video platforms [1]. 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 [2] and other false stories about presidential candidates [3]. Misinformation has also contributed to confusing the public about the reality of climate change [4, 5] and stoked skepticism about vaccine safety among the public [6, 7, 8]. 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 [9]. 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 [10].

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" [11]. In Europe, the European Commission has established a 'Code of Practice on Disinformation' [12] that enjoins platforms to voluntarily comply with a set of commitments [13]. 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 [14]. 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 [15]. 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 [16]. Facebook has developed the most extensive third-party fact-checking program with dozens of partner institutions to assist the company in this endeavour [17]. 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” [18]. 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 [19] and reducing false beliefs [20]. In an experimental setting, Pennycook et al. [21] 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 [22, 23]:

> *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 [16]. 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 analysed 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 [24].

A more recent article by Kornbluh et al. (2020) [25] 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 and colleagues [24]. 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 and colleagues [26] 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 [27]. The three studies mentioned above find rather different results due to different methodologies and use of sources that they labeled '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 [28]. 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 [29]. 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 directly from Facebook (using 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 Facebook 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 [30]. 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 [22, 23], 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, if a URL was shared in more than 1000 posts, we collected only the 1000 posts that received the highest number of interactions [31]; 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 [32] 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 Figure 1 for two examples).

The set of accounts analysed 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 [33].

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 Figures 2, 4, 5, 7 and 9) represents the 5% and 95% percentiles of the 1,000 medians.

3.2. Results

Figure 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’.

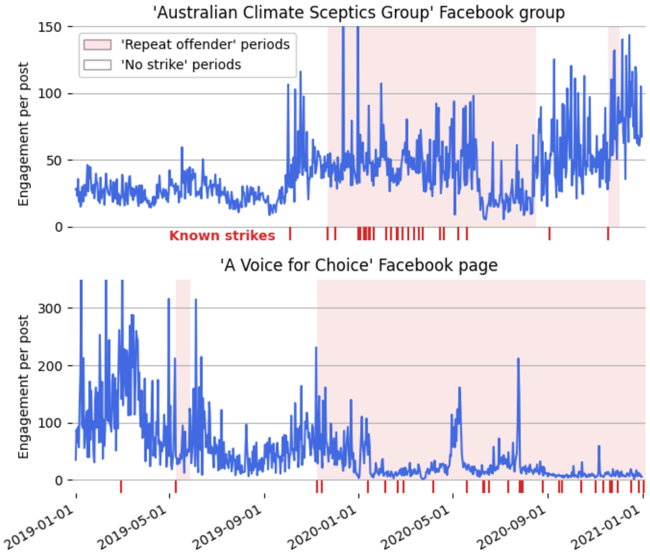


Figure 1: (Top panel) Average engagement (the sum of comments, shares, likes, ...) per post for the ‘Australian Climate Sceptics Group’ Facebook group for each day in 2019 and 2020. Each red line at the bottom represents the date of a known strike for this group according to Science Feedback data. The 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 Figure

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%.

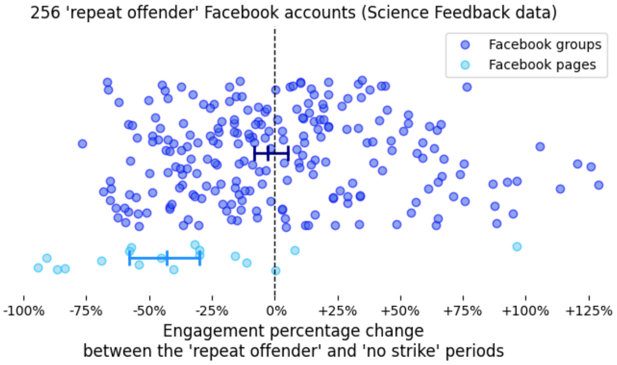


Figure 2: Percentage changes between the average engagement per post during the 'repeat offender' periods and the 'no strike' periods. Each deep blue dot represents a Facebook group, and each light blue dot a Facebook page. The bars show the medians for each set and their 90% confidence intervals (the 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.

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 (Figure 3). This metric, representing the total engagement generated by these accounts on Facebook (top panel), can be

[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 Figure 3).

decomposed as the number of posts published each day (middle panel) times the average number of engagement per post (bottom panel).

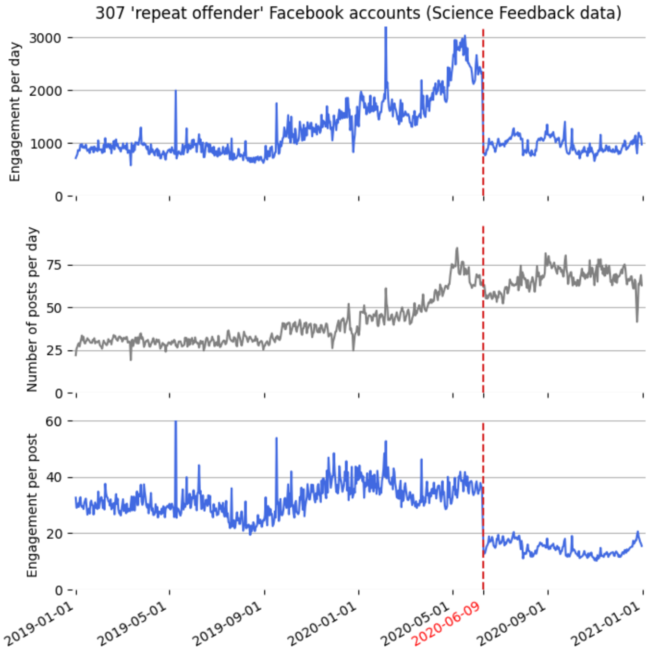


Figure 3: Metrics aggregated over the 307 Facebook accounts that repeatedly shared false news links identified using Science Feedback data. **(Top panel)** Total engagement per day averaged for all accounts. **(Middle panel)** Number of posts per day. **(Bottom panel)** Average engagement per post. The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

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 (Figure 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.

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 (Figure 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 = 4.6×10^{-17}).

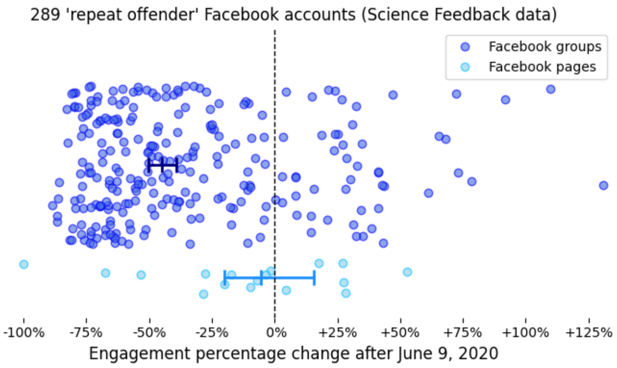


Figure 4: Percentage changes in the average engagement per post during a 30-day period before and after June 9, 2020. Each deep blue dot represents a Facebook group, and each light blue dot a Facebook page. The bars show the medians for each set and their 90% confidence intervals. The 289 'repeat offender' 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.

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 = 5.1×10^{-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

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

drop does not affect groups and pages equally, we reproduced Figure 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.

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)

4.1. Methods

We used data from the Social Science One organization [34], that builds partnerships between academia and private industries such as Facebook to share data and expertise. In July 2021, we had access to a new version of the Condor dataset [35], which contains all URLs shared publicly by at least 100 Facebook users between 2017 and 2021, as well as their fact-checking metadata. From this list, we extracted the 6,811 URLs that were shared in 2019 and 2020, that were fact-checked as 'False' and whose country in which it was shared most frequently was either the USA, Canada, Great Britain or Australia.

We then replicated as closely as possible the methods used in the previous section. Using CrowdTangle, we thus collected all the posts that shared one of the false links between January 1, 2019 and December 31, 2020, and focused on the 706 Facebook accounts (671 Facebook groups and 35 Facebook pages) that spread at least 24 false links. Then we used CrowdTangle again to collect all the posts published by those accounts in 2019 and 2020. Because the Condor dataset contained the date of the first fact-check done on a URL, we were able to infer the 'repeat offender' periods for each account and therefore conduct the same analysis as in the previous section.

Science Feedback being a third-party fact-checker working with Facebook, some of the URLs from Science Feedback 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 actually the same as the accounts analyzed previously. As the point of this new analysis is to replicate the previous results, we exclude the accounts whose engagement was already shown in the previous section. We thus show here the metrics for only the 503 'novel' accounts, which represent 476 groups and 27 pages (see Supplementary Figure S6).

4.2. Results

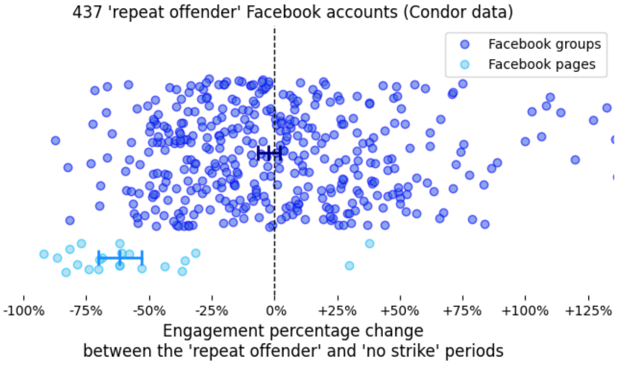


Figure 5: Same metric as on Figure 2 The 437 'repeat offender' accounts represented here were identified using Condor data, and have published at least one post during each period.

Our first objective is to verify that the repeat offender policy was applied only to Facebook pages, and not to groups during the 2019-2020 period. To do this, 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 Figure 5). The median percentage change is −5%, and the values are not significantly different from zero (W = 46495, p-value = 0.61).

The changes in engagement are also different for the groups and the pages (Figure 5). The percentage changes for the 414 Facebook groups are not different than zero (W = 41561, p-value = 0.57), with a median of −2%, while the values for the 23 Facebook pages are significantly different than zero (W = 29, p-value = 0.00041), and the median is −62%.

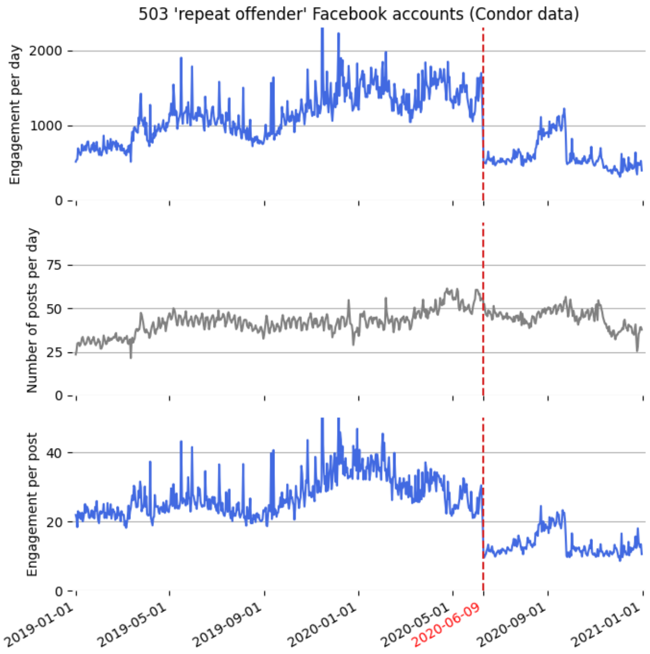


Figure 6: Same metrics as on Figure 3 aggregated over the 503 ‘repeat offender’ Facebook accounts identified using Condor data.

As in the previous section, we then analyzed the engagement received by the 503 repeat offenders accounts in 2019 and 2020 (see Figure 6). The ‘novel’ accounts replicated the slow rise in total engagement from September 2019 to June 2020, and the massive drop around June 9, 2020. Again, we observe that this measure set the engagement for ‘repeat offenders’ groups back to its early

2019 level.

The percentage change in engagement was then calculated for each account during a 30-day period before and after June 9, 2020 (Figure 7). The median percentage change is −26%, and 63% of the accounts experienced a decrease in engagement, the results being a little more modest than what was found previously. The values are still significantly different from zero (W = 42651, p-value = 3.8×10^{-5}).

When tested 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%. Also, when the engagement per post is plotted separately for groups and pages, we can see a drop in engagement only for groups (see Supplementary Figure S2).

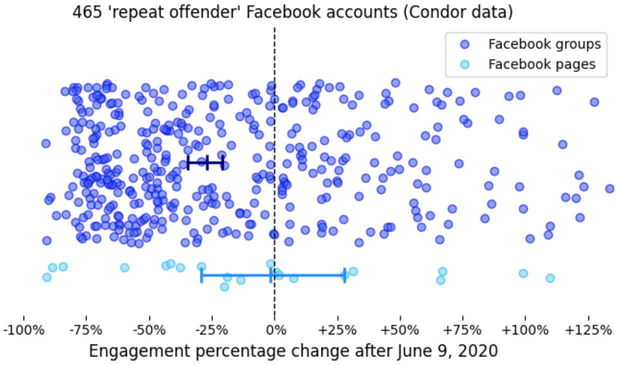


Figure 7: Same metric as on Figure 4. The 465 'repeat offender' accounts represented here were identified using Condor data, and have published at least one post one month before and one month after June 9, 2020.

To conclude, using a more complete dataset of 'False' URLs and collecting new Facebook accounts, we replicated our previous findings. Indeed we again find a sudden decrease in engagement for repeat offender Facebook groups in June 2020, and a decrease in engagement following the publication of two false links for repeat offender Facebook pages.

One limitation of the results is that this king of analysis is rather indirect, as we relied on the strike dates to infer the ‘repeat offender’ periods, and we cannot know for certain whether the pages investigated were actually under a ‘repeat offender’ status. For example, one could imagine that the ‘two strikes in less than 90 days’ rule may have changed over time, or that links fact-checked as ‘partly false’ or ‘missing context’ were also counted as strikes (only links fact-checked as ‘False’ were taken into account in our analysis). 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:

- ‘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 Figure 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.

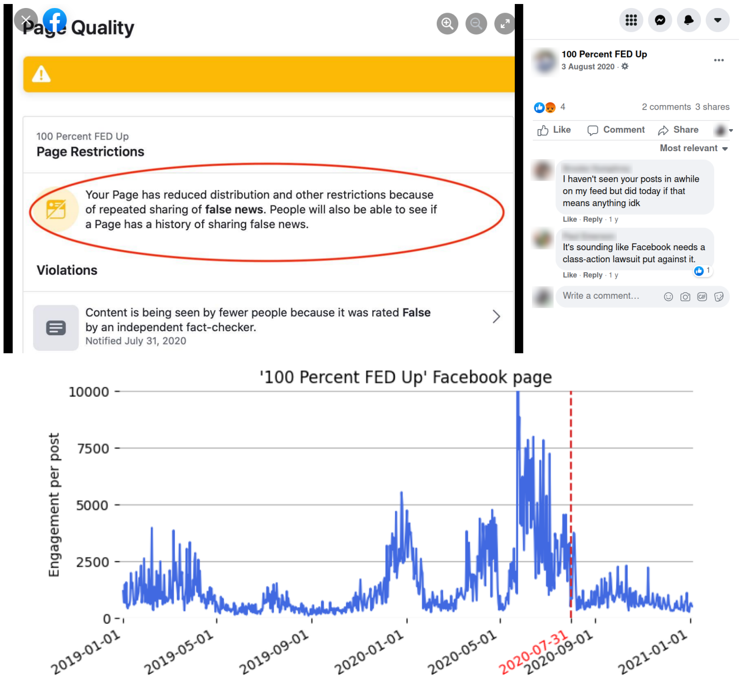
To verify whether Facebook applied any restriction to these pages, we collected all the posts that these 94 pages have published between January 1, 2019 and December 31, 2020 from the CrowdTangle API using the '/posts' endpoint. The collection was run on January 11, 2021. We were only able to collect 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 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 months 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

Figure 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 of that page over the past two years. We see a clear



Page Quality

100 Percent FED Up
Page Restrictions

Your Page has reduced distribution and other restrictions because of repeated sharing of **false news**. People will also be able to see if a Page has a history of sharing false news.

Violations

Content is being seen by fewer people because it was rated **False** by an independent fact-checker.
Notified July 31, 2020

100 Percent FED Up
3 August 2020 · ⚙

4 · 2 comments 3 shares

Like · Comment · Share

Most relevant

I haven't seen your posts in awhile on my feed but did today if that means anything idk
Like · Reply · 1 y

It's sounding like Facebook needs a class-action lawsuit put against it.
Like · Reply · 1 y

Write a comment…

Figure 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 infered from the date of the last violation on the screenshot ('Notified July 31, 2020').

decrease in engagement in August 2020, and when we compare the engagement during a 30-day period before and after July 31, 2020, the percentage change is −62%.

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 Figure 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). We can thus suggest that the 'reduced distribution' status is associated with a modest decrease in engagement.

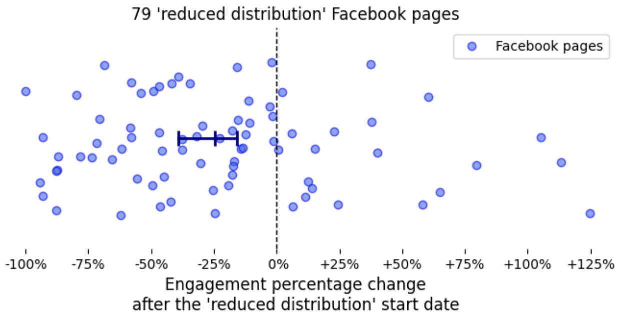


Figure 9: Percentage changes in average engagement per post during a 30-day period before and after the reduced distribution start date. Each dot represents a Facebook page. The bars show the median and its 90% confidence interval. The 79 'reduced distribution' pages represented here were identified because they shared a 'reduced distribution' notification from Facebook in 2020.

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 (also see Supplementary Figure S3 to observe the engagement dynamics). This confirms that Facebook pages have most likely not been affected by the sudden *reduce* measure implemented in June 2020 and evidenced in the previous sections.

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 'repeat offenders', which are accounts that repeatedly share false information. However, the effects of the platforms' diverse policies to tackle misinformation remains understudied [36]. The present research article aims to contribute to filling this knowledge gap by verifying the application and measuring 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 [32]. 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 no evidence that this policy is 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 novel accounts sharing misinformation. We again observed a significant decrease (median of -62%) in engagement for 'repeat offender' pages, while the engagement for 'repeat offender' groups remained stable.

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. 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. On one hand, to identify perfectly repeat offender periods from a list of False URLs, we should have access to all the False URLs that this given accout has shared, and the two URL data sources we relied on are both more or less incomplete. Moreover the 'two strikes in 90 days' rule used to infer repeat offender periods may not be the exact rule used by Facebook, or the rule may have changed with time. On the other hand, the date at which a page shared their 'reduced distribution' notification may be months later after the page received the notification.

There is no public data indicating when a group or page is under a repeat offender status, as this information is a private notification sent to Facebook to the concerned accounts only. The only way to monitor the repeat offender policy was thus to deduce which groups or pages should be reduced from the data available to us. We found consistent results in the three different analyses, with a median decrease in engagement ranging from -62% to -24%, and it hints that these different methods estimated more or less correctly the repeat offender periods. Furthermore there might be a lag between the strike date and the application of the 'repeat offender' policy, to allow for human verification. If so, it should be of a few days and have minimal impact on our calculations using 30 days or more windows. Because of these potential biases, it is likely that the true size effect of the 'repeat offender' interventions is underestimated.

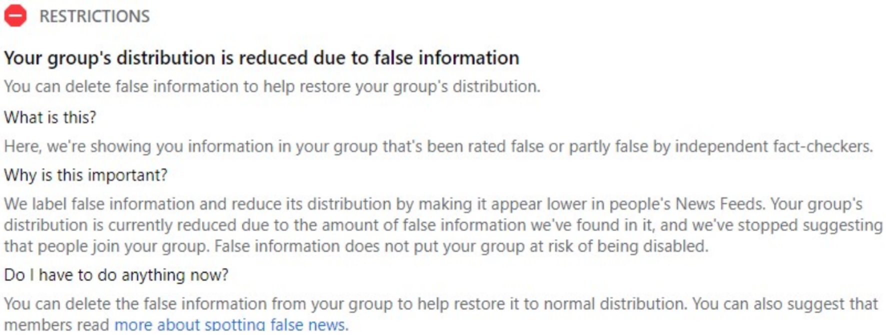
Although we observe a global reduction in engagement, there is a large heterogeneity across the different 'repeat offender' pages (see Figures 2, 5 and 9). The engagement of some popular pages have actually increased, such as the 'Tucker Carlson Tonight' page with a 38% increase (from 104k to 143k interactions per post) following the 'reduced distribution' notification from Facebook. The engagement of the 'Mark Levin' page remained rather stable after the notification (change of 2%), going from 20.3k to 20.7k interactions per post. It is possible that these popular pages had their distribution already reduced for months before sharing the notification, or that they compensated the decrease in engagement led by the reduce intervention by a simultaneous gain of followers, but a recent article points toward an alternative explanation. Some high-profile Facebook users such as celebrities, politicians and journalists might be exempted from the normal enforcement processes, according to company documents revealed by the Wall Street Journal [37].

By analyzing the time series of the repeat offenders' engagement over the past two years, we also discovered a sudden drop 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 second 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 repeat offenders’ groups, in ways that differ from its stated policy against repeat offenders pages. 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 [25, 26]. 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 [38]. 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 Figures 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 [39]. 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 fighting misinformation is a relatively new issue for platforms that appeared with the 2016 American elections, and the misinformation regulations are constantly changing. It appears that Facebook is now also applying its reduce policy on misinformation groups, as one of the 'repeat offender' group analyzed has shared in 2021 a 'reduced distribution' notification from Facebook (see Figure 10).



RESTRICTIONS

Your group's distribution is reduced due to false information

You can delete false information to help restore your group's distribution.

What is this?

Here, we're showing you information in your group that's been rated false or partly false by independent fact-checkers.

Why is this important?

We label false information and reduce its distribution by making it appear lower in people's News Feeds. Your group's distribution is currently reduced due to the amount of false information we've found in it, and we've stopped suggesting that people join your group. False information does not put your group at risk of being disabled.

Do I have to do anything now?

You can delete the false information from your group to help restore it to normal distribution. You can also suggest that members read more about spotting false news.

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

The 'reduced distribution' notification is different for groups and pages. Notably groups are informed by Facebook that: "You can delete false information to help restore your group's distribution" (see Figure 10). In contrast, page owners cannot get rid of a strike in the same way as group owners: "Note that deleting a post will not eliminate the strike against the Page or domain", although they can correct the posts and submit an appeal to fact-checkers for the strike to be lifted [40]. As the followers of a group, and not just its administrators, can post in the group, group content can be hard to control. Maybe that is why this exception to the 'repeat offender' restrictions was created for groups. We would nevertheless argue that it makes the policy easier to be circumvented for groups repeatedly sharing misinformation. Furthermore, Facebook has announced in May 2021 that an individual's Facebook account will also be reduced if they repeatedly share misinformation content [41]. It would thus be interesting to replicate our findings on the 2021 engagement data to monitor the effects of these new measures.

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 [42]. As a consequence, researchers [19, 43] and journalists [44, 45] 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 [46]),

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.

Acknowledgements

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 Programme d'Investissements d'Avenir (ANR-19-MPGA-0005).

Author contributions

Emmanuel M. Vincent: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing - Review & Editing. **Héloïse Théro:** Data Curation, Investigation, Software, Validation, Visualization, Writing - Original Draft.

References

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

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

[3] H. Allcott, M. Gentzkow, Social media and fake news in the 2016 election, Journal of economic perspectives 31 (2) (2017) 211–36. `doi:10.1257/jep.31.2.211`.

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

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

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

[7] M. Lahouati, A. De Coucy, J. Sarlangue, C. Cazanave, Spread of vaccine hesitancy in france: What about youtube™?, Vaccine 38 (36) (2020) 5779–5782. doi:10.1016/j.vaccine.2020.07.002.

[8] F. Pierri, B. Perry, M. R. DeVerna, K.-C. Yang, A. Flammini, F. Menczer, J. Bryden, The impact of online misinformation on us covid-19 vaccinations, arXiv preprint arXiv:2104.10635.

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

[10] Parse.ly's network referrer dashboard, https://www.parse.ly/resources/data-studies/referrer-dashboard, accessed on 2021-07-08.

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

[12] Code of practice on disinformation, https://ec.europa.eu/digital-single-market/en/code-practice-disinformation, [European Commission] (2021).

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

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

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

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

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

[18] Rating options for fact-checkers, https://www.facebook.com/business/help/341102040382165, [Facebook Help].

[19] P. Mena, Cleaning up social media: The effect of warning labels on likelihood of sharing false news on facebook, Policy & internet 12 (2) (2020) 165–183. doi:10.1002/poi3.214.

[20] E. Porter, T. J. Wood, 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). doi:10.1073/pnas.2104235118.

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

[22] Fact-checking on facebook, https://www.facebook.com/business/help/2593586717571940, [Facebook Help].

[23] Facebook's enforcement of fact-checker ratings, https://www.facebook.com/business/help/297022994952764, [Facebook Help].

[24] H. Allcott, M. Gentzkow, C. Yu, Trends in the diffusion of misinformation on social media, Research & Politics 6 (2) (2019) 2053168019848554. doi:10.1177/2053168019848554.

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

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

[27] Iffy quotient, https://csmr.umich.edu/projects/iffy-quotient/.

[28] 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.

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

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

[31] https://github.com/CrowdTangle/API/wiki/Links.

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

[33] F. Wilcoxon, Individual comparisons by ranking methods, in: Breakthroughs in statistics, Springer, 1992, pp. 196–202. doi:10.2307/3001968.

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

[35] S. Messing, C. DeGregorio, B. Hillenbrand, G. King, S. Mahanti, Z. Mukerjee, C. Nayak, N. Persily, B. State, A. Wilkins, Facebook privacy-protected full urls data set, [data set], Havard Dataverse, V7 (2021). doi:10.7910/DVN/TDOAPG.

[36] I. V. Pasquetto, B. Swire-Thompson, M. A. Amazeen, F. Benevenuto, N. M. Brashier, R. M. Bond, L. C. Bozarth, C. Budak, U. K. Ecker, L. K. Fazio, et al., Tackling misinformation: What researchers could do with social media data, the Harvard Kennedy School Misinformation Review (2020). doi:10.37016/mr-2020-49.

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

[38] N. F. Johnson, N. Velásquez, N. J. Restrepo, R. Leahy, N. Gabriel, S. El Oud, M. Zheng, P. Manrique, S. Wuchty, Y. Lupu, The online competition between pro-and anti-vaccination views, Nature 582 (7811) (2020) 230–233. doi:10.1038/s41586-020-2281-1.

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

[40] Issue a correction or dispute a rating, https://www.facebook.com/business/help/997484867366026, [Facebook Help].

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

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

[43] W. Yaqub, O. Kakhidze, M. L. Brockman, N. Memon, S. Patil, Effects of credibility indicators on social media news sharing intent, in: Proceedings of the 2020 chi conference on human factors in computing systems, 2020, pp. 1–14. doi:10.1145/3313831.3376213.

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

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

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

[47] Coverage of the coronavirus on web and social, https://go.newswhip.com/2020_03_Covid-19_LP.html, [NewsWhip] (2020).

[48] A. Davey, Facebook sent flawed data to misinformation researchers, https://www.nytimes.com/live/2020/2020-election-misinformation-distortions#facebook-sent-flawed-data-to-misinformation-researchers, [The New York Times] (2021).

SUPPLEMENTARY INFORMATION

Engagement dynamics plotted separately for groups and pages

In this article, we find a clear difference in how Facebook groups and pages are regulated by Facebook, which explains why the data is often plotted separately for these two kinds of accounts. The only exceptions are the engagement dynamics shown in Figures 3 and 6, in which all the ‘repeat offender’ accounts - groups and pages - are reprensented together. However, the June drop is only affecting Facebook groups and not pages, but this difference is not visible on the above mentioned figures representing all the accounts.

This is why we have plotted here the engagement per posts separately for groups and pages for the accounts identified using the Science Feedback dataset (see Figure S1) and for the accounts identified using the Condor dataset (see Figure S2). We can observe that the engagement per post do remain stable for the ‘repeat offender’ pages in June 2020.

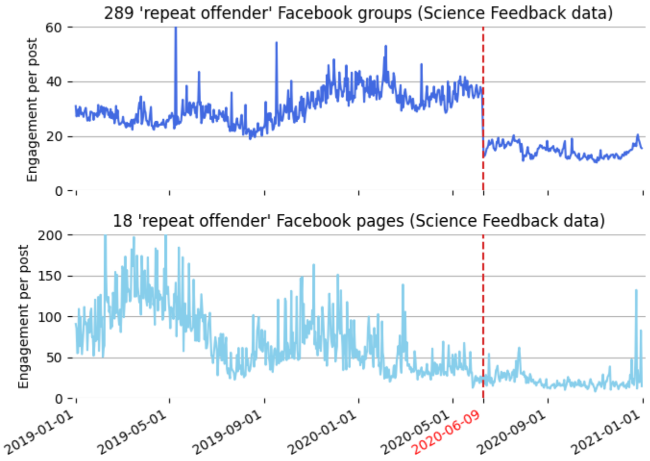


Figure S1: Average engagement per post in 2019-2020 plotted separately for the 289 groups (top panel) and the 18 pages (bottom panel) identified as ‘repeat offender’ using Science Feedback data.

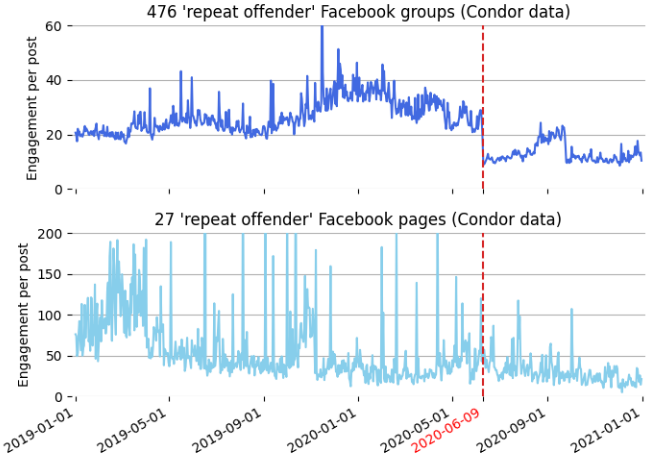


Figure S2: Average engagement per post in 2019-2020 plotted separately for the 476 groups (top panel) and the 27 pages (bottom panel) identified as 'repeat offender' using Condor data.

We also plotted here the engagement dynamics for the set of misinformation pages that shared a 'reduced distribution' notification (see Figure S3). As with the previous sets of pages shown above, there is no reduction in engagement for 745 these pages in June 2020.

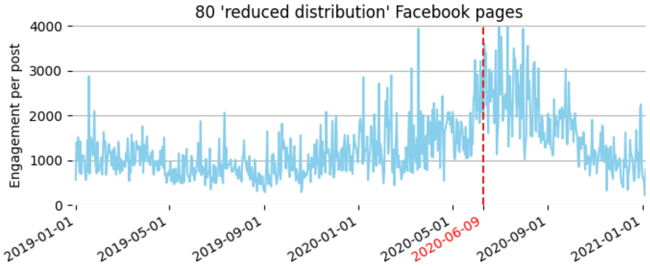


Figure S3: Average engagement per post in 2019-2020 for the 80 'reduced distribution' pages, identified because they shared a 'reduced distribution' notification from Facebook.

These graphs illustrate that only Facebook groups, and not pages, were affected by the reduce measure implemented on June 9, 2020.

Engagement dynamics in 2019-2020 for a control set of accounts

We compared the dynamics of the ‘repeat offender’ accounts to those of a control set of accounts, which consisted of Facebook pages and groups associated with established news outlets that we expect to have received no false fact-checks on their posts.

To identify such a set, we used a report from NewsWhip [47] that identified the 10 media outlets that communicated the most during the early phase of the COVID-19 pandemic (first half of 2020), i.e., NBC, The Daily Mail, CNN, Fox News, The Independent, BBC, The New York Times, The Washington Post, Yahoo and The New York Post. We searched the outlets’ names on Facebook and created a list of 10 pages and six groups that displayed a verified ‘blue check’. We also searched for more groups as they are the accounts supposed to be affected by the June drop. We added to this set 19 Facebook groups created before June 2020 that are either associated with a media outlet (such as the ‘Brexit latest - the Independent’ group) or that have a science focus (such as the ‘WE ARE SCIENTISTS’ group). Using CrowdTangle, we collected all the posts published by these accounts between January 1, 2019 and December 31, 2020.

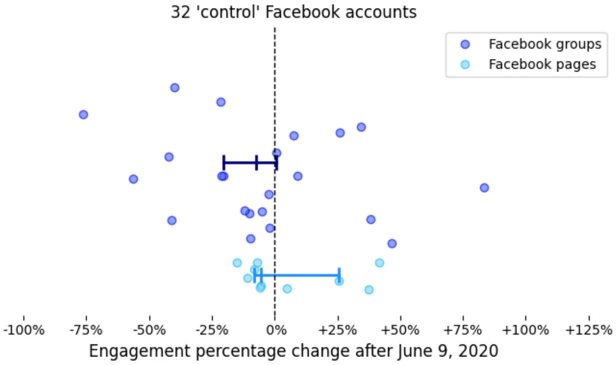


Figure S4: Same metrics as on Figure 4 aggregated over the ‘control’ Facebook accounts that published at least once during 30 days before and after June 9, 2020.

The percentage changes after June 9, 2020 are not significantly different from zero for groups (W = 95, p-value = 0.32, median = −7%) and for pages (W = 27, p-value = 1, median = −6%, see Figure S4). Therefore, contrary to what we observe for the 'repeat offender' groups, we found no drop in engagement in June 2020 for the 'control' groups. This observation further supports the hypothesis that the drop observed for the 'repeat offender' groups is specifically targeted at these misinformation groups, and not a feature that broadly affected Facebook groups.

Overlap between the lists of false URLs

In the two first methods, we used two different sources to get a list of False URLs fact-checked in 2019-2020, but there should be an overlap between these two lists. Indeed, Science Feedback is a third-party fact-checker partnering with Facebook [30], and the URLs fact-checked by Science Feedback were transfered to Facebook. We can thus imagine that the list of URLs from Science Feedback would be included in the list from Condor.

However the only URLs that are in Condor are the ones shared by more than 100 users on Facebook, which excludes the less viral URLs fact-checked by Science Feedback. Moreover, as Condor is one of the largest social science research dataset ever constructed, issues related to data quality, validity and fidelity are expected to be found [35]. For example it was recently revealed that the engagement data in Condor was only based on around half of the U.S. users and thus incomplete, because the views of the users that were not politically classified were not taken into account [48]. Although this error should not impact the list of URLs we used in this article, other issues might have altered the list of False URLs, and that reason could also explain why some URLs from Science Feedback were excluded from the Condor list.

To compare the two lists of URLs, we first normalized all the URLs with the same method, and then built a Venn diagram from the two lists of normalized URLs (see Figure S5). The overlap was found to be smaller than expected. Indeed only 32% of the URLs in Science Feedback were also in Condor, sug-

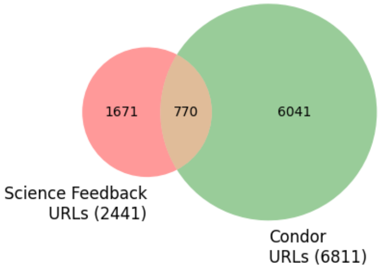


Figure S5: Overlap between the list of False URLs from Science Feedback and the list of False URLs from Condor.

gesting that most of the URLs fact-checked were shared by less than 100 users. Furthermore, the URLs from Science Feedback represented only 11% of all the URLs in Condor. As Science Feedback is only one of the 60+ fact-checking organizations partnering with Facebook [17], we can see that its fact-checked URLs are actually well represented in the Condor dataset.

Overlap between the different sets of accounts analyzed

As we used different methodologies to obtain three different lists of 'repeat offender' accounts, we verify how much of these accounts were redundant in the different lists.

In the third analysis only pages were found, and thus we only compare the lists of Facebook groups collected between the first and second analyses (see Figure S6 left panel). Although the lists of false URLs do not overlap that much between the two data sources, we found that two third (67%) of the groups identified using Science Feedback data were also obtained from Condor data. This can be explained because the ULRs in common between Science Feedback and Condor were the most viral ones (shared by more than 100 users) and thus these URLs played an important role to identify accounts repeatedly sharing misinformation.

We also compare the lists of pages found using the three different methods

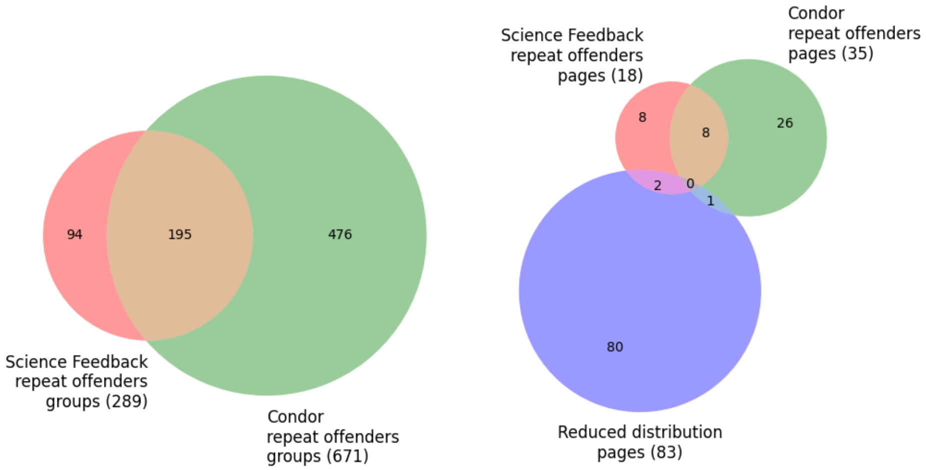


Figure S6: (Left panel) Overlap between the two lists of Facebook groups identified using Science Feedback data (first analysis) and Condor data (second analysis). **(Right panel)** Overlap between the three lists of Facebook pages identified using Science Feedback data (first analysis), using Condor data (second analysis) and by sharing a 'reduced distribution' notification (third analysis).

(Figure S6 right panel). We again found a significant overlap between the page list for two first analyses, as 8 pages out of the 18 pages identified using Science Feedback data were also found using Condor data. Interestingly the overlap between of the two first lists and the 83 pages sharing a 'reduced distribution' notification was almost null (with only 2 pages in common with the first list, and 1 page with the second list). The 'reduced distribution' notifications used to identified the last set of pages were mostly shared during the last semester of 2020. Because we kept only pages that shared 24 or more False URLs in the two first analyses, it is possible that this method was biased to select pages that rather received their 'reduced distribution' notification in 2019 or earlier in 2020. This would explain the negligible overlap between pages identified as having shared False URLs and pages identified as having shared a 'reduced distribution' notification. Obtaining such different sets of pages confirm the interest of using both methods to investigate the effects of the 'repeat offender' policy.