Investigating Facebook's actions 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 efficiency of this measure. First, combining data from a fact-checking organization and CrowdTangle, we identified a set of public accounts (groups and pages) that have shared misinformation repeatedly during the 2019-2020 period. While we observe a decrease in engagement for pages after they shared two or more 'false news', such a reduction is not observed for 'repeat offender' groups. This result was replicated using a more complete set of misinformation URLs shared by Facebook (Condor data). We also 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. Finally we identified a set of pages claiming to be put 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%. Our results highlight easy steps Facebook could take to reduce misinformation, such as to enforce their 'repeat offender' policy more forcefully on pages, and to start applying it to groups.

*Corresponding authors.

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

Keywords: Misinformation, Content moderation, Algorithmic transparency, 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 confused the public about the reality of climate change [4, 5] and stoked skepticism about vaccine safety among the public [6, 7]. 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 [8]. 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 [9].

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" [10]. In Europe, the European Commission has established a 'Code of Practice on Disinformation' [11] that enjoins platforms to voluntarily comply with a set of commitments [12]. 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 [13]. Facebook has developed the most extensive third-party fact-checking program with dozens of partner institutions to assist the company in this endeavour [14]. Facebook informs page or group owners when published posts on their pages or groups are marked as misinformation, inviting them to correct the posts. Facebook also states that the virality of the posts marked as 'False' or 'Partly False' will be reduced.

Facebook's *reduce* policy is not only applied to individual posts, but also to organizations and communities that often publish posts containing misinformation, as indicated by this statement in their publishers' help center [15, 16]:

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

Facebook ranks each post by assigning to it a relevancy score, where a high score leads to a high likelihood of the post to appear on a user's newsfeed. Doing so, Facebook can make a post or a whole account less visible by decreasing the relevancy score of its content; this is precisely the *reduce* measure [13].

So far Facebook has not provided data showing how their *reduce* policy is implemented, which would allow researchers to quantify its impact on the spread of misinformation. To the best of our knowledge, the impact of the *reduce* policy has not yet been audited directly. Hence the present research article departs from articles studying the overall levels of misinformation on platforms [17, 18, 19], by focusing on monitoring a specific policy against misinformation.

We used CrowdTangle, a public insights tool owned and operated by Face-

book, to access Facebook data [20]. 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 [21]. To investigate the effect of the reduce policy, we thus analyzed the engagement per post received by 'repeat offender' accounts. Indeed, one would expect this metric to decline if the accounts' posts become less visible in Facebook's feed.

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 more complete source of misinformation URLs obtained directly from Facebook (Condor data). Using another method to identify 'repeat offender' accounts, we finally studied the engagement metrics of a set of Facebook pages claiming to be under reduced distribution.

Research questions.

- Was the policy aiming to reduce the distribution of misinformation repeat offenders actually enforced by Facebook during the 2019-2020 period?
- If so, what was the magnitude of the reduction applied? And is there a difference between Facebook groups and Facebook pages?
- Does this action have an impact on the spread of misinformation on Facebook, i.e., can we see a global increase or decrease in engagement for the repeat offender accounts through time?

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

We used data from Science Feedback, which is part of Facebook's third-party fact-checking program [22]. Science Feedback is a fact-checking organization, where academics review the credibility of science-related claims and articles.

2.1. Methods

We obtained from Science Feedback a list of 4,000+ URLs reviewed by its team. 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 fact-checked as 'Altered' also contribute to the 'repeat offender' status [15, 16], but we found no such rating in the Science Feedback data. The list of 'false news links' was obtained on January 4, 2021 and cover links flagged in 2019 and 2020.

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 [23]. We focused on the accounts that spread misinformation the most often, choosing a threshold of 24 different false news links shared over the past two years.

The corresponding 307 Facebook accounts (289 Facebook groups and 18 Facebook pages) are referred to as 'repeat offenders accounts'. 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, an undisclosed source obtained by a journalist [24] indicated 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 an example).

2.2. Results

Figure 1 displays the engagement metrics for one 'repeat offender' group named 'Australian Climate Sceptics Group'. 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 (see Figure 1). 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. However we observe a 61% increase in engagement.

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 results are different. For the 238 Facebook groups, the percentage changes are not significantly different

[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, the second semester of 2020 was excluded for its vastly diminished and not representative engagement level (see Figure 3).

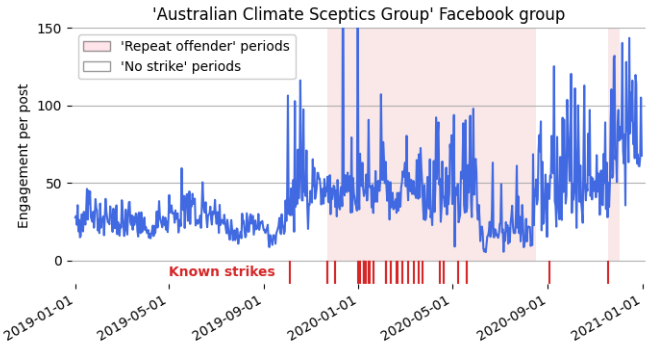

Figure 1: Average engagement (the sum of comments, shares, likes, ...) per post for the 'Australian Climate Sceptics Group' Facebook group for each day in 2019 and 2020. Each red line at the bottom represents the date of a known strike for this group according to the Science Feedback data. The areas shaded in red represent the 'repeat offender' periods as defined by the 'two strikes in 90 days' rule.

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%.

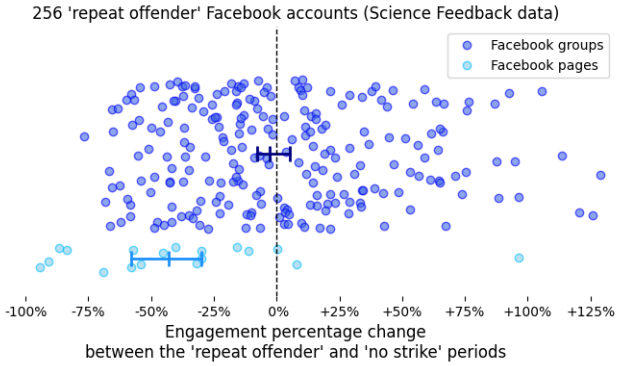

Figure 2: Percentage changes between the average engagement per post during the 'repeat offender' periods and the 'no strike' periods. Each deep blue dot represents a Facebook group, and each light blue dot a Facebook page. The bars show the medians for each set and their 90% confidence intervals (the intervals are estimated using a bootstrap method). The 256 'repeat offender' accounts represented here were identified by the 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 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 (Figure 3). This drop has cut the groups' engagement per post in half, but it was compensated by the fact that the overall activity of 'repeat offender' has doubled between 2019 and 2020. The engagement for 'repeat offender' groups was thus reset by this intervention to its pre-pandemic level.

To further quantify this 'June drop', we calculated the percentage change in engagement 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 $= 4.6 \times 10^{-17}$).

Again the results differ between Facebook pages and Facebook groups. While the percentage changes for the 271 groups are significantly different than zero ($W = 7599$, p-value $= 5.1 \times 10^{-17}$), with a median of -45%, the 18 pages appears to be not affected by the decrease ($W = 73$, p-value $= 0.61$), with a

[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).

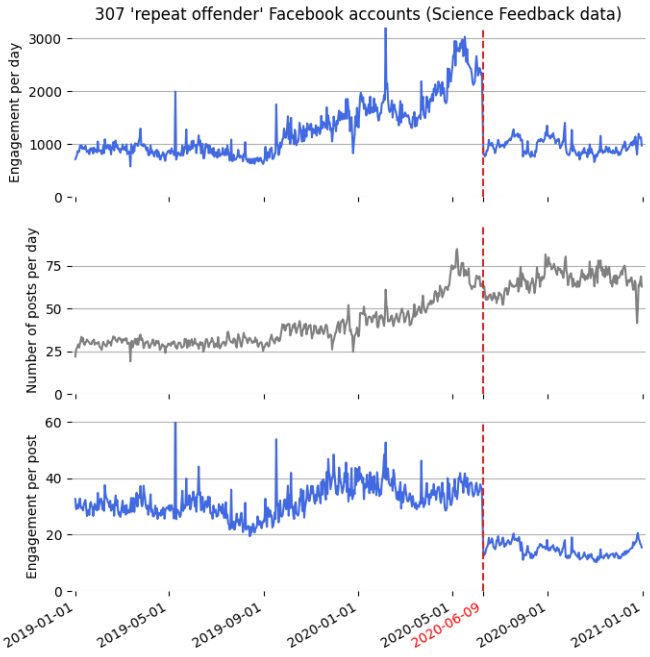

Figure 3: (Top panel) Average engagement per day. **(Middle panel)** Number of posts per day. **(Bottom panel)** Average engagement per post. The dotted red line marks the date of June 9, 2020, when a sudden drop in engagement is observed. The metrics were aggregated over the 307 'repeat offender' Facebook accounts identified by the Science Feedback data.

median percentage change of only -5%.

To verify whether this drop was specific to this set of groups, we compared these dynamics to those of a control set of accounts consisting of Facebook pages and groups associated with established news outlets that did not publish misinformation. No such drop in total or per post engagement metrics was observed around June 9, 2020 (see Supplementary Figure X).

We can only explain such a massive change by a modification in how Facebook's 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 only took action against these groups via this one-shot measure in June.

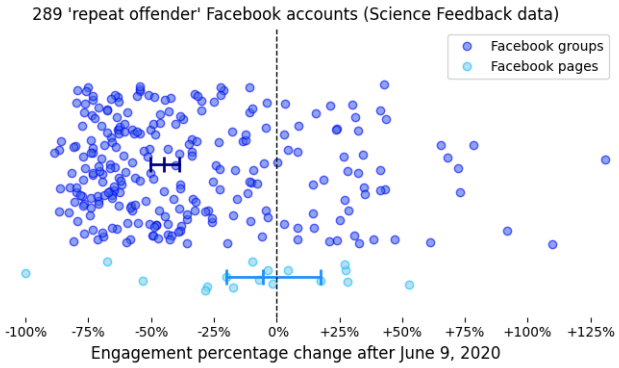

Figure 4: Percentage changes in the average engagement per post during a 30-day period before and after June 9, 2020. Each deep blue dot represents a Facebook group, and each light blue dot a Facebook page. The bars show the medians for each set and their 90% confidence intervals. The 289 'repeat offender' accounts represented here were identified by the Science Feedback data, and have published at least one post one month before and one month after June 9, 2020.

One limitation of the results described in this section is that we obtained the links labelled as 'False' from only one fact-checking organization (Science Feedback), while Facebook partners with over 60 fact-checking organizations [14]. The true 'repeat offender' periods could thus be longer than the ones inferred, potentially changing the magnitude of the 'reduce' effect.

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

3.1. Methods

We used data from the Social Science One organization [25], 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 [26], 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, most of the URLs from Science Feedback were also contained in the Condor dataset (see Supplementary Figure X). Thus an important part of the 'repeat offender' groups and pages obtained from the Condor URLs were actually the same as the accounts analyzed previously (see Supplementary Figure X). The point of this new analysis was to replicate the previous results with a more complete URL dataset and for this reason, we excluded the accounts whose engagement was already shown in the previous section. We thus show here the results for 503 'novel' accounts: 476 groups and 27 pages.

3.2. Results

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

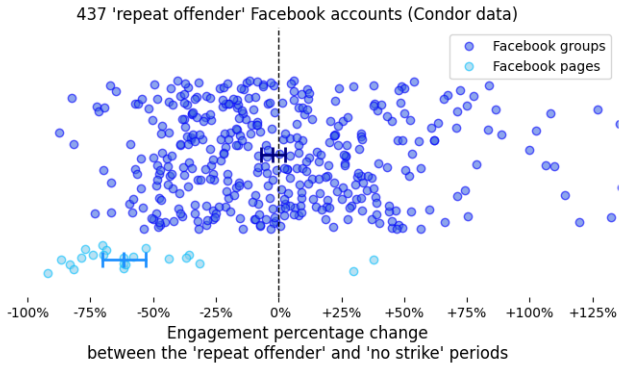

Figure 5: Percentage changes between the average engagement per post during the 'repeat offender' periods and the 'no strike' periods. Each deep blue dot represents a Facebook group, and each light blue dot a Facebook page. The bars show the medians for each set and their 90% confidence intervals. The 437 'repeat offender' accounts represented here were identified by the Condor data, and have published at least one post during each period.

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%.

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

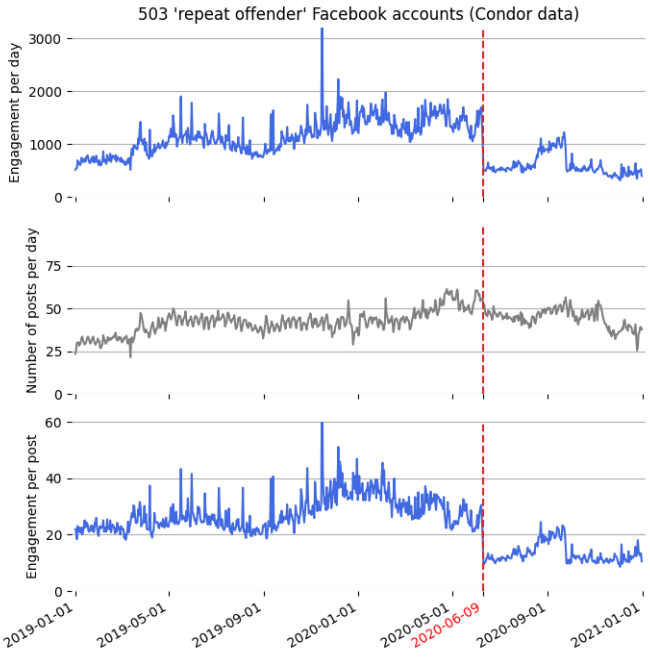

Figure 6: (Top panel) Average engagement per day. **(Middle panel)** Number of posts per day. **(Bottom panel)** Average engagement per post. The dotted red line marks the date of June 9, 2020, when a sudden drop in engagement is observed. The metrics were aggregated over the 503 'repeat offender' Facebook accounts identified by the Condor data.

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%.

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

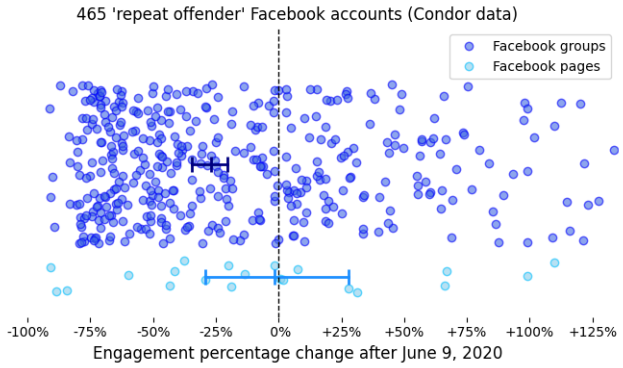

Figure 7: Percentage changes in the average engagement per post during a 30-day period before and after June 9, 2020. Each deep blue dot represents a Facebook group, and each light blue dot a Facebook page. The bars show the medians for each set and their 90% confidence intervals. The 465 'repeat offender' accounts represented here were identified by the Condor data, and have published at least one post one month before and one month after June 9, 2020.

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.

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

4.1. 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'.

The date of the last 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.

4.2. Results

Figure 8 shows a screenshot of the Facebook notification shared by the 'I Love Carbon Dioxide' page on April 28, 2020, and the average engagement per post of that page over the past two years. The engagement does not appear to be reduced after April 28, 2020. When we compare the engagement during a 30-day period before and after this date, the percentage change is 2%, indicating that the engagement is not affected by the 'repeat offender' status.

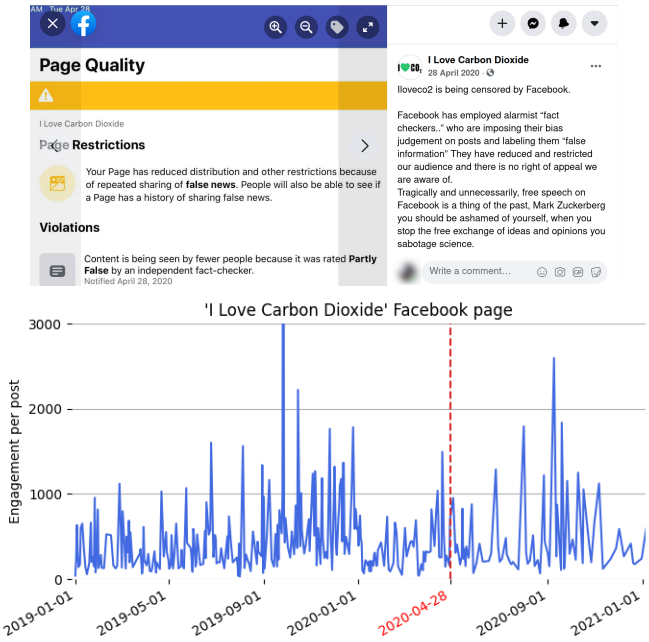

Figure 8: (Top panel) Screenshot of a post from the 'I Love Carbon Dioxide' Facebook page sharing a 'reduced distribution' notification from Facebook. **(Bottom panel)** Average engagement per post for the "I Love Carbon Dioxide" page for each day in 2019 and 2020. The dotted red line represents the reduced distribution start date.

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 82 Facebook pages that published at least one post during each period (see Figure 9). The median percentage change is -24%, and a Wilcoxon test reveals that the percentage changes are significantly different from zero ($W = 911$, p-value $= 0.00026$). We can thus suggest that the 'reduced distribution' status is associated with a modest decrease in engagement.

However, there is a large heterogeneity across the different Facebook pages. The engagement of some popular pages have actually increased following the notification, such as the 'Tucker Carlson Tonight' page with a 38% increase (from 104k to 143k interactions per post).

Finally, we verify whether an important drop in engagement also occurred

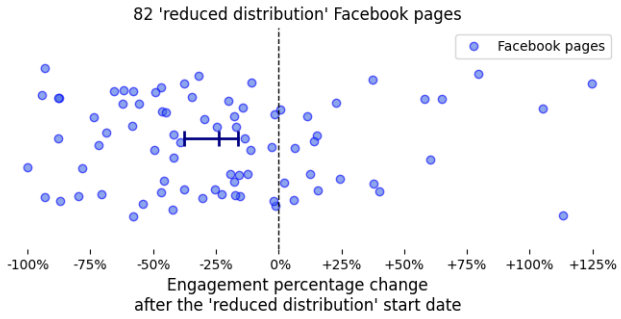

Figure 9: Percentage changes in average engagement per post during a 30-day period before and after the reduced distribution start date. Each dot represents a Facebook page. The bars show the median and its 90% confidence interval. The 82 'reduced distribution' pages represented here were identified because they shared a 'reduced distribution' notification from Facebook in 2020.

in June 2020 for this set of Facebook pages. When we compare the engagement metrics before and after June 9, 2020, the percentage changes are not significantly different from zero (W = 1093, p-value = 0.055), and the median percentage change is 3%. This confirms that Facebook pages have most likely not been affected by the *reduce* measure implemented on June 9, 2020 and evidenced in the previous sections.

5. 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 [27]. 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 the reach of 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. 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.

ADD CONDOR RESULTS

As a second 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 -24%), 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.

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%. The 18 Facebook pages from the first sample, as well as the 83 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.

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 [28]. 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 (add proportions?). In the interest of curbing the spread of misinformation, applying its 'repeat offender' policy to groups as well as to pages would help Facebook decrease the amount of misinformation in their users' feeds.

It is also 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 [18, 19]. 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 [29]. 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 Figure 3 top panel).

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 [30]. As a consequence, researchers [31, 32] and journalists [33, 34] 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 and 3) a small number of accounts is responsible for a large proportion of the misinformation being shared (at least regarding COVID-19 [35]), 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.

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 vaccina-

tion 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] 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).

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

[10] 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).

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

[12] 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.

[13] 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).

[14] 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).

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

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

[17] 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.

[18] 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-ru [GMF The German Marshall Fund of the United States] (2020).

[19] 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).

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

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

[22] 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).

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

[24] 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).

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

[26] 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 (2020). doi:10.7910/DVN/TDOAPG.

[27] 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.

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

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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).

[34] 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).

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