Investigating Facebook's actions against accounts that repeatedly share misinformation

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

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

Abstract

SHORTEN AND ADD NEW CONDOR RESULTS

Like many web platforms, Facebook is under pressure to regulate misinformation. According to the company, users that repeatedly share misinformation ('repeat offenders') will have their distribution reduced, but little is known about the implementation or the 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. While we observe a decrease in engagement for pages (median of -43%) after they shared two or more 'false news', such a reduction is not observed for groups. However, we discover that groups have been affected in a different way with a sudden drop in their average engagement per post that occurred around June 9, 2020. No public information was given by Facebook about this sudden decrease. This drop has cut the groups' engagement per post in half, but it was compensated by the fact that the overall activity of 'repeat offenders' has doubled between 2019 and 2020. Second, we identified pages that have been warned by Facebook and have shared a screenshot of the 'reduced distribution' notification they have received. We found that their engagement per post following the notification decreased by a modest amount (median of -24%), with some popular pages actually gaining more engagement. Our results highlight

*Corresponding authors.

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

easy steps Facebook could take to reduce misinformation, such as to enforce their 'repeat offenders' policy more forcefully on pages, and to start applying it to groups.

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

1. Introduction

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)

To investigate the effect of fact-checking on Facebook accounts that repeatedly share misinformation, we first analyzed the engagement per post received by these accounts. One would expect this metric to decline if the accounts' posts become less visible in Facebook's feed.

2.1. Methods

We used data from Science Feedback, which is part of Facebook's third-party fact-checking program [1]. Science Feedback is a fact-checking organization, where academics review the credibility of science-related claims and articles. 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. 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 [2]. 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 offenders' 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 offenders' is not specified. However, an undisclosed source obtained by a journalist 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.* [3]

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.

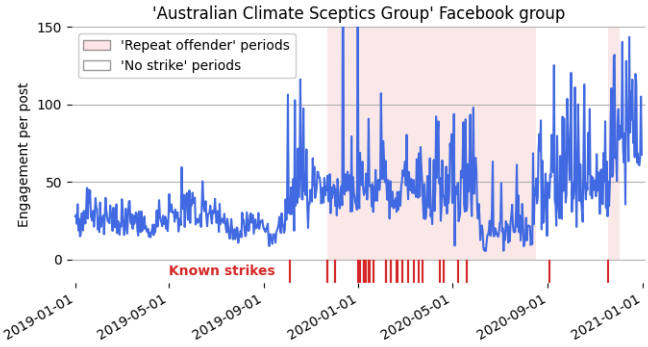

Figure 1: Average engagement (the sum of comments, shares, likes, ...) per post for the 'Australian Climate Sceptics Group' Facebook group for each day in 2019 and 2020. Each red line at the bottom represents the date of a known strike for this group, and the areas shaded in red represent the 'repeat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

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 average percentage change is 7%, and the median −6%. 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. The median percentage change for Facebook groups is −3% (not significantly different from zero), while the median for Facebook pages is −43%. A Wilcoxon test applied only to the Facebook pages' percentage changes, shows they are significantly different from zero (W = 21, p-value = 0.0034).

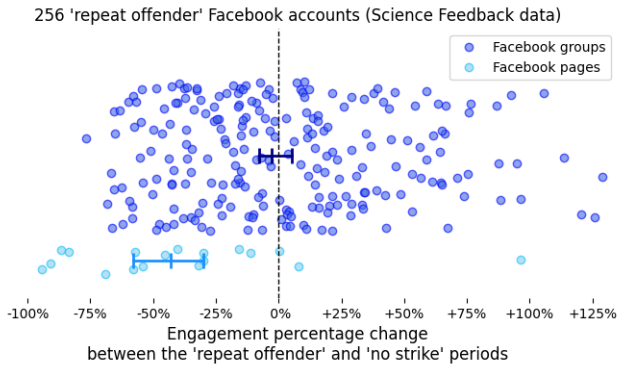

Figure 2: Percentage changes between the average engagement per post during the 'repeat offender' periods and the 'no strike' periods. Each deep blue dot represents a Facebook group, and each light blue dot a Facebook page. The bars show the medians for each set and their 90% confidence intervals. Confidence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

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

[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).

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).

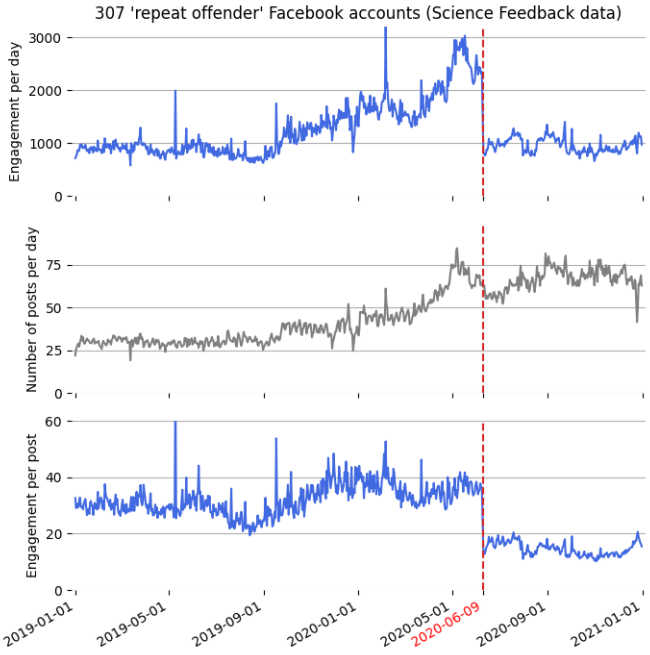

Figure 3: Metrics aggregated over the 307 Facebook accounts that repeatedly shared false news links. **(Top panel)** Average engagement per day. **(Middle panel)** Number of posts per day. **(Bottom panel)** Average engagement per post. The dotted red line marks the date of June 9, 2020, when a sudden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

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 average percentage change is -21%, and the median -43%.

Most of the accounts (219 out of 289) experienced a decrease in engagement[2], and a Wilcoxon test indicates that these percentage changes are significantly different from zero (W = 9012, p-value = 4.6×10^{-17}).

It appears that the Facebook pages are not affected by this decrease, with a median percentage change of −5%, while the groups have a median percentage change of −45%. When tested separately, the Facebook pages' percentage changes are not significantly different from zero (W = 73, p-value = 0.61).

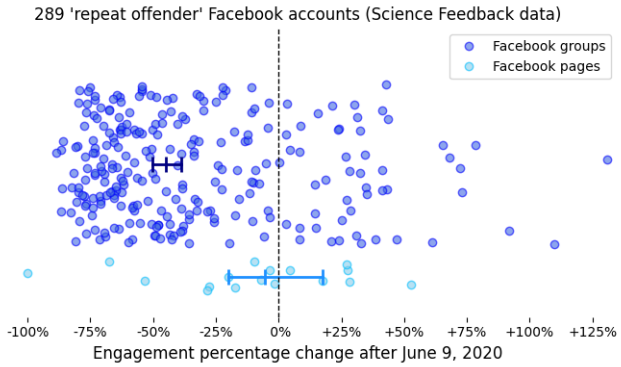

Figure 4: Percentage changes in the average engagement per post during a 30-day period before and after June 9, 2020. Each deep blue dot represents a Facebook group, and each light blue dot a Facebook page. The bars show the medians for each set and their 90% confidence intervals.

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.

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 de-

[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).

crease in engagement for 'repeat offenders' pages, we observed no such link for 'repeat offenders' groups. Hence it seems that Facebook only took action against these groups via this one-shot measure in June.

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 [4]. 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 [5], 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 [6], 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 found to be different for the groups and the pages. The median percentage change for the 414 Facebook groups is -2%, while the median for the 23 Facebook pages is -62% (Figure 5). A Wilcoxon test applied only to the percentage changes of the Facebook pages shows they are significantly different from zero ($W = 29$, p-value $= 0.00041$).

ADD FIGURE

EXPAND THIS SECTION 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.

ADD FIGURE

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}). The 23 pages have a median percentage change of -2% (not significantly different from zero: W = 133, p-value = 0.89), while the 442 groups have a median percentage change of -27%.

ADD FIGURE

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.

References

[1] 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).

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

[3] 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).

[4] 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).

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

[6] 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.