Persistence of online misinformation: Investigating Facebook's actions against "repeat offenders"

Anonymous TTO submission

Abstract

Like many web platforms, Facebook is under pressure to regulate misinformation. According to the company, pages that repeatedly share misinformation ("repeat offenders") will have their distribution reduced, but little is known about the implementation or the efficacy of this measure. First, combining data from a fact-checking organization and CrowdTangle, we did not find that Facebook groups repeatedly sharing misinformation had reduced engagement metrics, but a drastic drop was observed around June 9, 2020. No public information was given by Facebook about this sudden decrease. Second, we collected the posts from pages that shared a 'reduced distribution' notification from Facebook, and found a decrease in engagement following the notification. Facebook's reduced distribution policy against repeat offenders thus appears to be applied only on Facebook pages, and not on Facebook groups.

1 Introduction

The general public is increasingly getting news related information online, through search engines, social media and video platforms (Mitchell et al., 2016). Hence the spread of misinformation through these platforms has recently received growing attention. Recent studies, along with the political context of January 2021 in the United States, show how the presence of misinformation online can contribute to negative societal consequences. Namely it can fuel false beliefs, such as the idea of a massive voter fraud during the US 2020 presidential election, which may have led to the January 6, 2021 insurrection at the U.S. Capitol (Benkler et al., 2020) and other false stories about presidential candidates (Allcott and Gentzkow, 2017). Misinformation has also confused the public about the reality of climate change (Brulle, 2018; Porter et al., 2019) and stoked skepticism about vaccine safety among the public (Featherstone and Zhang, 2020; Lahouati et al., 2020). In April 2020, a questionnaire from the Reuters Institute found that people in the UK use online sources more often than offline sources when looking for information about the coronavirus. Among social media platforms, Facebook was the most widely used with 24% of the respondents saying they used Facebook to access COVID-19 information in the last seven days (Fletcher et al., 2020). The structural importance of Facebook to the media landscape is confirmed by Parse.ly's dashboard, showing that the visitors to their 2500+ online media sites are referred by Facebook in 25% of the cases[1].

Lawmakers and regulators are increasingly pressuring platforms to limit the spread of disinformation. In the US, the House of Representatives organized hearings and convened representatives of the main platforms to testify on how they are being weaponized to spread "misinformation and conspiracy theories online" (Donovan et al., 2020). In Europe, the European Commission has established a 'Code of Practice on Disinformation'[2] that enjoins platforms to voluntarily comply with a set of commitments (Heldt, 2019). However, there is little data available and few established processes to monitor the implementation of these measures and quantify their actual impact. This is what we propose to tackle in this paper by offering a methodology to monitor Facebook's implementation of its policy to reduce the distribution of accounts repeatedly spreading misinformation. We chose to focus on Facebook as it is the biggest social media platform with more than 2 billion users worldwide.

[1] https://www.parse.ly/resources/data-studies/referrer-dashboard, accessed on 2021-07-08.

[2] https://ec.europa.eu/digital-single-market/en/code-practice-disinformation.

Facebook announced a three-part policy to fight against 'misleading or harmful content': they claim to *remove* harmful information, *reduce* the spread of misinformation and *inform* people with additional context[3]. Facebook has developed the most extensive third-party fact-checking program with dozens of partner institution to assist the company in this endeavour[4]. 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[5]:

> *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[6].

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 (Allcott et al., 2019; Kornbluh et al., 2020; Resnick et al., 2018), by focusing on monitoring a specific policy against misinformation.

We used CrowdTangle, a public insights tool owned and operated by Facebook, to access Facebook data (Team, 2021). CrowdTangle exclusively tracks public content, and provides access to engagement metrics (such as number of likes, shares and comments), but not to the reach (number of views) of content[7]. We first investigated how Facebook enforces its 'reduce' policy by combining data from one of Facebook's fact-checking partners identifying URLs sharing misinformation and tracking engagement metrics of the Facebook accounts that repeatedly share such misinformation. We further investigated the effects of Facebook's policy on engagement metrics of a set of Facebook pages claiming to be under reduced distribution.

2 Investigating the 'reduce' policy on Facebook groups repeatedly sharing misinformation

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[8]. Science Feedback is a fact-checking organization, in which 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 to True' 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 gathered 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

[3] https://about.fb.com/news/2018/05/inside-feed-reduce-remove-inform/

[4] https://www.facebook.com/business/help/341102040382165

[5] https://www.facebook.com/business/help/2593586717571940, https://www.facebook.com/business/help/297022994952764

[6] https://about.fb.com/news/2018/05/inside-feed-reduce-remove-inform/

[7] https://help.crowdtangle.com/en/articles/3192685-citing-crowdtangle-data, https://help.crowdtangle.com/en/articles/4558716-understanding-and-citing-crowdtangle-data

[8] https://sciencefeedback.co/science-feedback-partnering-with-facebook-in-fight-against-misinformation/

more than 1000 posts, we collected only the 1000 posts that received the highest number of interactions[9]. We focused on the accounts that spread the most misinformation, and chose 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 named '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. 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"[10].

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 sharing a misinformation link was published after the corresponding fact-check, we used the date of the post as the strike date. If the account first shared a link, which was then fact-checked as 'False', the fact-check publication 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 are shown as red lines at the bottom, and the inferred 'repeat offender' periods are shaded in red. The average engagement per post varies throughout the

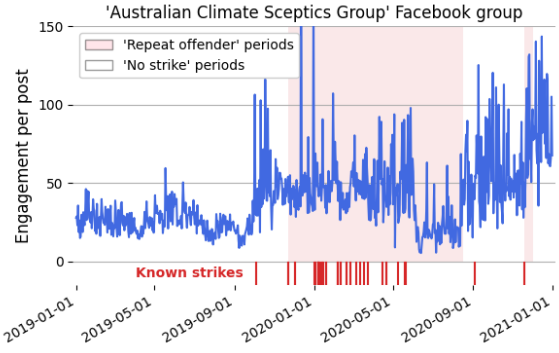

Figure 1: Average engagement (the sum of comments, shares, likes, ...) per post for the 'Australian Climate Sceptics Group' Facebook group for each day in 2019 and 2020. Each red line at the bottom represents the date of a known strike for this group, and the areas shaded in red represent the 'repeat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

measuring period, but not appears to be related with the shift of 'repeat offender' and 'no strike' periods (see Figure 1). When comparing the average engagement metrics between the 'repeat offender' and the 'no strike' periods, we observe a 61% increase in engagement during the 'repeat offender' periods.

To generalize, we calculate the percentage change between the 'repeat offender' and the 'no strike' periods for each of the 256 Facebook accounts that has published at least one post during each period (see Figure 2) [11]. When considering all accounts, 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 considering groups and pages separately, it appears that 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%. When applied only to the Facebook pages' percentage changes, the Wilcoxon test shows they are significantly different from zero ($W = 21$, p-value $= 0.0034$).

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

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

[10] https://www.nbcnews.com/tech/tech-news/sensitive-claims-bias-facebook-relaxed-misinformation-rules-conservative-pages-n1236182

[11] 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 for its vastly diminished and not representative engagement level (see Figure 3).

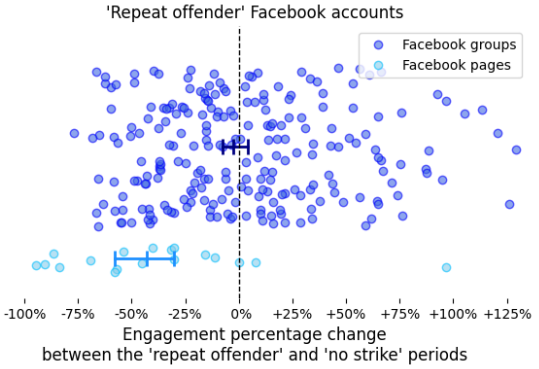

Figure 2: Percentage changes between the average engagement per post during the 'repeat offender' periods and the 'no strike' periods. Each deep blue dot represents a Facebook group, and each light blue dot a Facebook page. The bars show the median and its 90% confidence interval. Confidence intervals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

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).

We observe a rise in total engagement per day from September 2019 to June 2020, while it remained stable from January to September 2019. This rise is explained by the increase in activity of the misinformation accounts, with a doubling of the number of posts per day between September 2019 and June 2020, while the engagement per post remained about constant. Around June 9, 2020, the total engagement metrics suddenly experience a massive drop that is entirely explained by a corresponding drop in engagement per post (Figure 3).

To further quantify this 'June drop', we calculated the percentage change in engagement for each account between 30 days before June 9, 2020 and 30 days after (Figure 4). When considering all accounts, the average percentage change is -21%, and the median -43%. Most of the accounts (219 out of 289) experienced a decrease in engagement[12]. A Wilcoxon test indicates that these percentage changes are significantly different from zero (W = 9012, p-value = 4.6×10^{-17}).

[12] 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).

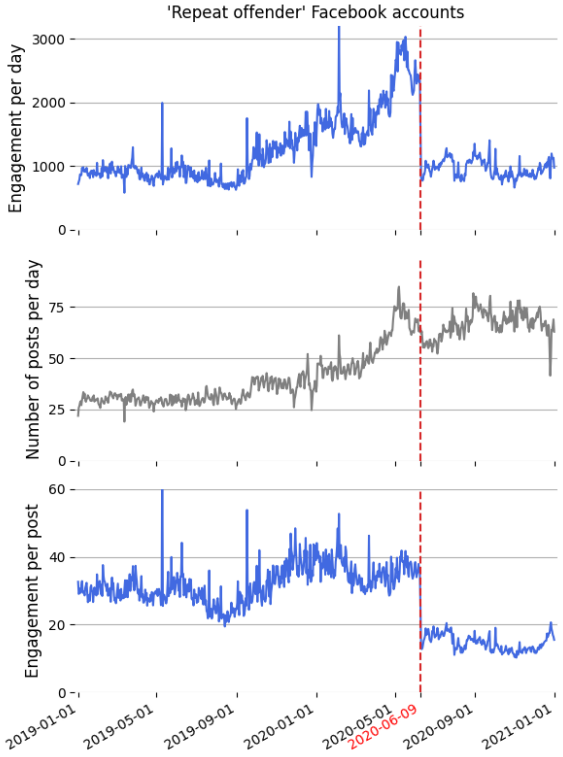

Figure 3: Metrics aggregated over the 307 Facebook accounts that repeatedly shared false news links. **(Top panel)** Engagement per day. **(Middle panel)** Number of posts per day. **(Bottom panel)** Average engagement per post. The dotted line marks the date of June 9, 2020, when a sudden drop in engagement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

When considering groups and pages separately, 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).

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. Despite this one-shot measure, the results showed a global lack of correlation between the fact-checks' dates and a decrease in engagement for 'repeat offenders' Facebook groups. However we only took into account the links labelled as 'False' by one fact-checking organization (Science Feedback), while Facebook partners with over 60 fact-checking organizations[13]. The true repeat offender periods could thus be longer

[13] https://about.fb.com/news/2020/04/covid-19-misinfo-update/

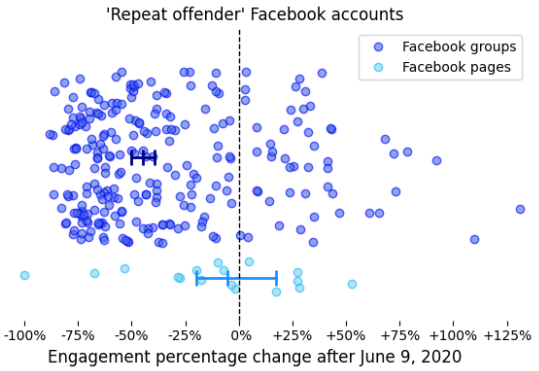

Figure 4: Percentage changes between the average engagement per post 30 days before June 9, 2020 and 30 days after. Each deep blue dot represents a Facebook group, and each light blue dot a Facebook page. The bars show the median and its 90% confidence interval.

than the ones inferred here. Moreover, although we observed no effect on Facebook groups, Facebook pages did appear to have a reduced engagement per post during the 'repeat offender' periods.

3 Investigating the 'reduce' policy on self-declared 'repeat offenders' Facebook pages

In the present section we focused on Facebook pages, and used a different methodology to collect accounts which we are certain are under 'repeat offender' status.

3.1 Methods

We analyzed a set of pages that have publicly shared a message claiming to be placed under 'repeat offender' status and including a screenshot as evidence, often to express their indignation about it. We had observed this behaviour in two popular pages ('Mark Levin' and '100 Percent FED Up') and searched for more such examples. To assemble a list of self-declared repeat offenders, we searched Facebook via the CrowdTangle API, using the '/posts/search' endpoint of the API on November 25, 2020, for posts published since January 1, 2020 with the following keywords:

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

We manually opened the hundreds of resulting posts, and kept the posts we found to meet the following criteria (see Figure 5 Top panel for an example):

- they shared a screenshot of the Facebook message received
- the Facebook message indicated 'Your page has reduced distribution and other restrictions because of repeatedly sharing of false news.'
- the page name was visible on the screenshot message

We obtained a list of 94 pages associated with one of these posts. 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 violation notification was used as the inferred start date of reduced distribution when the date was visible in the screenshot. When the screenshot did not include the date of the last violation notification, we used the date of the post as the inferred start date of reduced distribution.

3.2 Results

Figure 5 shows the reduced distribution 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 between 30 days before and 30 days after April 28, 2020, the percentage change is 2%, indicating this metric is not influenced by the 'repeat offender' notification.

To generalize, we calculate the percentage change in engagement between 30 days after and before the reduced distribution start date for each of the 82 Facebook pages that published at least one post during each period (see Figure 6). The average percentage change was −16%, the median

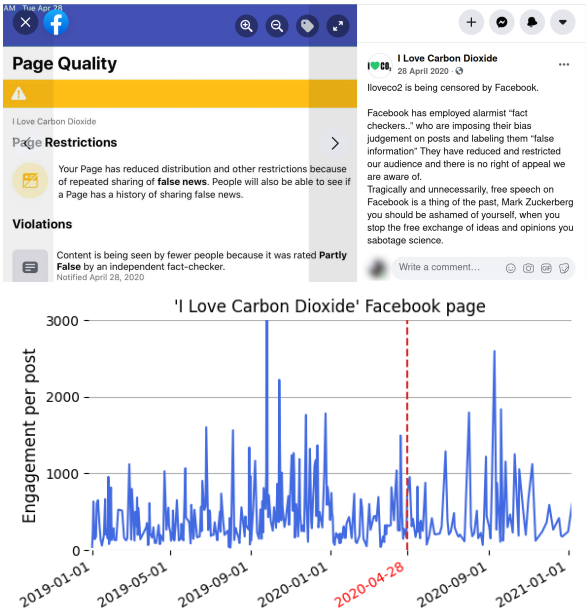

Figure 5: (Top panel) Screenshot of a post from the 'I Love Carbon Dioxide' Facebook page sharing a reduced distribution notification from Facebook. **(Bottom panel)** Average engagement per post for the "I Love Carbon Dioxide" page for each day in 2019 and 2020, with the reduced distribution start date shown in red.

-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' notification is generating a modest decrease in engagement. However, there is a wide range of variations across the different Facebook pages, with some large pages actually increasing their engagement following the notification, such as the 'Tucker Carlson Tonight' page with a 38% increase (from 104k to 143k interactions per post).

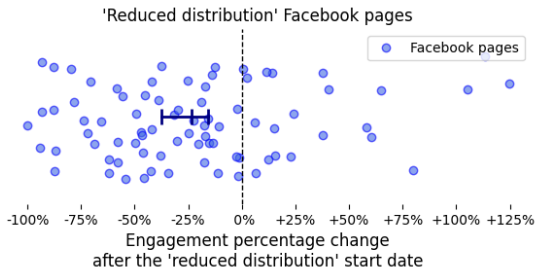

Figure 6: Percentage changes between the average engagement per post 30 days before the 'reduced distribution' start date and 30 days after. Each dot represents a Facebook page. The bars show the median and its 90% confidence interval.

Finally, we verify whether an important drop in engagement also occurred in June 2020 on this set of Facebook pages, which are different from the ones in Section 2. When we compare the average engagement metrics between 30 days after and 30 days before 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 probably not been affected by the *reduce* measure implemented on June 9, 2020 and evidenced in the previous section.

4 Discussion

Facebook, the most widely used social media platform in the world, has announced a series of measures to curb the spread of misinformation, namely by reducing the visibility of 'repeat offenders', which are accounts that repeatedly share false information. However, the effects of platforms' diverse policies to tackle misinformation remains understudied (Pasquetto et al., 2020). The present research article aims to contribute to filling this knowledge gap by 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 of two false news links over a three-month period is supposed to be penalized by reducing the distribution of the account's content. However, we find no compelling evidence that this policy of Facebook is actually leading to a significant reduction in engagement for Facebook groups, which would be an expected outcome of reducing the distribution. But a decrease in engagement was observed for the 18 Facebook pages in our sample. As a second step, we gathered 83 Facebook pages that shared a notification they received from Facebook announcing that their account was under reduced distribution. The pages' engagement metrics were significantly lower after the notification, suggesting that the 'reduced distribution' measure was indeed applied to the pages. It should be noted that no group was found when searching for accounts sharing a reduced distribution notification, which reinforces the hypothesis that the 'repeat offender' policy is applied only on Facebook pages, and not on groups.

By analyzing the evolution of the repeat offenders' engagement, we also discovered a sudden drop affecting Facebook groups around June 9, 2020. For many groups, the decrease was drastic (around 70% - 80%). Interestingly, the 18 Facebook pages from the first sample were not concerned by this decrease, as the 83 pages from the second sample. This 'June drop' does not correspond to any official communication by Facebook on that matter. It indicates that Facebook has very likely taken internal decisions that heavily impact the organic reach of repeat offenders' groups, in ways that differ from its stated policy. Such finding makes us wonder whether similar *reduce* measures were suddenly implemented at different times and on other lists of accounts. 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 aims: pages are supposed to be for official communication, while groups should foster interactions between users[14]. Pages thus have to be public, while groups can be public or private. Moreover pages' posts can be monetized and promoted. Despite these differences, both pages and groups can spread false news, and in the interest of curbing the spread of misinformation, it is hard to understand why the 'repeat offender' policy was implemented only on pages. Similarly, it is not clear why only repeat offender Facebook groups, and not pages, were reduced in June 2020. Studies have highlighted that misinformation persists at high levels on Facebook and other platforms (Kornbluh et al., 2020; Resnick et al., 2018). It is possible that such pressure has driven Facebook to suddenly regulate the engagement of groups speading misinformation, as a 'quick fix' to compensate for the absence of a repeat offender policy. In a context where the overall engagement for these misinformation groups was rising, the massive decrease on June 9, 2020 has succeeded in bringing the engagement level back to its early 2019 values (see Figure 3).

Online misinformation can be a threat to societies around the world, and the role of platforms in its regulation has been the subject of intense debate over the past few years (Rogers, 2020; De Gregorio and Stremlau, 2020). As a consequence, researchers (Mena, 2020; Yaqub et al., 2020) and journalists[15] have begun to monitor the platforms' actions against misinformation and their efficacy. Investigating Facebook's policy against 'repeat offenders', we observed a striking difference of treatment between the pages and groups that is not clearly stated in Facebook's communication. We emphatize the need for further research to thoroughly verify and shed light on the platforms' actions against misinformation.

[14] https://www.facebook.com/help/337881706729661/

References

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–36.

Hunt Allcott, Matthew Gentzkow, and Chuan Yu. 2019. Trends in the diffusion of misinformation on social media. *Research & Politics*, 6(2):2053168019848554.

Yochai Benkler, Casey Tilton, Bruce Etling, Hal Roberts, Justin Clark, Robert Faris, Jonas Kaiser, and Carolyn Schmitt. 2020. Mail-in voter fraud: Anatomy of a disinformation campaign. *Available at SSRN*.

R Brulle. 2018. 30 years ago global warming became front-page news–and both republicans and democrats took it seriously. *The Conversation*.

Giovanni De Gregorio and Nicole Stremlau. 2020. Internet shutdowns in africa— internet shutdowns and the limits of law. *International Journal of Communication*, 14:20.

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

Jieyu Ding Featherstone and Jingwen Zhang. 2020. Feeling angry: the effects of vaccine misinformation and refutational messages on negative emotions and vaccination attitude. *Journal of Health Communication*, 25(9):692–702.

Richard Fletcher, Antonis Kalogeropoulos, Felix M Simon, and Rasmus Kleis Nielsen. 2020. Information inequality in the uk coronavirus communications crisis. *Reuters Institute for the Study of Journalism*.

[15] https://www.economist.com/graphic-detail/2020/09/10/facebook-offers-a-distorted-view-of-american-news, https://www.nytimes.com/2020/11/24/technology/facebook-election-misinformation.html

Amélie Heldt. 2019. Let's meet halfway: Sharing new responsibilities in a digital age. *Journal of Information Policy*, 9:336–369.

K Kornbluh, A Goldstein, and E Weiner. 2020. New study by digital new deal finds engagement with deceptive outlets higher on facebook today than run-up to 2016 election. gmf the german marshall fund of the united states. viitattu 16.12. 2020.

Marin Lahouati, Antoine De Coucy, Jean Sarlangue, and Charles Cazanave. 2020. Spread of vaccine hesitancy in france: What about youtube™? *Vaccine*, 38(36):5779–5782.

Paul Mena. 2020. Cleaning up social media: The effect of warning labels on likelihood of sharing false news on facebook. *Policy & internet*, 12(2):165–183.

Amy Mitchell, Jeffrey Gottfried, Michael Barthel, and Elisa Shearer. 2016. The modern news consumer: News attitudes and practices in the digital era. *Pew Research Center*.

Irene V Pasquetto, Briony Swire-Thompson, Michelle A Amazeen, Fabrício Benevenuto, Nadia M Brashier, Robert M Bond, Lia C Bozarth, Ceren Budak, Ullrich KH Ecker, Lisa K Fazio, et al. 2020. Tackling misinformation: What researchers could do with social media data. *The Harvard Kennedy School Misinformation Review*.

Ethan Porter, Thomas J Wood, and Babak Bahador. 2019. Can presidential misinformation on climate change be corrected? evidence from internet and phone experiments. *Research & Politics*, 6(3):2053168019864784.

Paul Resnick, Aviv Ovadya, and Garlin Gilchrist. 2018. Iffy quotient: A platform health metric for misinformation. *Center for Social Media Responsibility*, 17.

Richard Rogers. 2020. Deplatforming: Following extreme internet celebrities to telegram and alternative social media. *European Journal of Communication*, 35(3):213–229.

CrowdTangle Team. 2021. Crowdtangle. *Facebook, Menlo Park, California, United States*.

Waheeb Yaqub, Otari Kakhidze, Morgan L Brockman, Nasir Memon, and Sameer Patil. 2020. Effects of credibility indicators on social media news sharing intent. In *Proceedings of the 2020 chi conference on human factors in computing systems*, pages 1–14.